Pedestrian Mobility Mining with Movement Patterns

Dissertation

zur
Erlangung des Doktorgrades (Dr. rer. nat.)

der
Mathematisch-Naturwissenschaftlichen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

Thomas Liebig

aus

Chemnitz

Bonn, 2013

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Gutachter:	Prof. Dr. Stefan Wrobel
2. Gutachter:	Prof. Dr. Armin B. Cremers
Tag der Promotion:	19.07.2013
Erscheinungsjahr:	2013

Thomas Liebig

Technical University of Dortmund
Department of Computer Science VIII

and

University of Bonn
Department of Computer Science III

and

Fraunhofer Institute for Intelligent Analysis
and Information Systems IAIS

Schloss Birlinghoven
53754 Sankt Augustin
Germany

`thomas.liebig@tu-dortmund.de`

Abstract

In street-based mobility mining, pedestrian volume estimation receives increasing attention, as it provides important applications such as billboard evaluation, attraction ranking and emergency support systems. In practice, empirical measurements are sparse due to budget limitations and constrained mounting options. Therefore, estimation of pedestrian quantity is required to perform pedestrian mobility analysis at unobserved locations. Accurate pedestrian mobility analysis is difficult to achieve due to the non-random path selection of individual pedestrians (resulting from motivated movement behaviour), causing the pedestrian volumes to distribute non-uniformly among the traffic network. Existing approaches (pedestrian simulations and data mining methods) are hard to adjust to sensor measurements or require more expensive input data (e.g. high fidelity floor plans or total number of pedestrians in the site) and are thus unfeasible.

In order to achieve a mobility model that encodes pedestrian volumes accurately, we propose two methods under the regression framework which overcome the limitations of existing methods. Namely, these two methods incorporate not just topological information and episodic sensor readings, but also prior knowledge on movement preferences and movement patterns.

The first one is based on Least Squares Regression (LSR). The advantage of this method is the easy inclusion of route choice heuristics and robustness towards contradicting measurements. The second method is Gaussian Process Regression (GPR). The advantages of this method are the possibilities to include expert knowledge on pedestrian movement and to estimate the uncertainty in predicting the unknown frequencies. Furthermore the kernel matrix of the pedestrian frequencies returned by the method supports sensor placement decisions. Major benefits of the regression approach are (1) seamless integration of expert data and (2) simple reproduction of sensor measurements. Further advantages are (3) invariance of the results against traffic network homeomorphism and (4) the computational complexity depends not on the number of modeled pedestrians but on the traffic network complexity.

We compare our novel approaches to state-of-the-art pedestrian simulation (Generalized Centrifugal Force Model) as well as existing Data Mining methods for traffic volume estimation (Spatial k-Nearest Neighbour) and commonly used graph kernels for the Gaussian Process Regression (Squared Exponential, Regularized Laplacian and Diffusion Kernel) in terms of prediction performance (measured with mean absolute error). Our methods showed significantly lower error rates.

Since pattern knowledge is not easy to obtain, we present algorithms for pattern acquisition and analysis from *Episodic Movement Data*. The proposed analysis of *Episodic Movement Data* involve spatio-temporal aggregation of visits and flows, cluster analyses and dependency models.

For pedestrian mobility data collection we further developed and successfully applied the recently evolved Bluetooth tracking technology. The introduced methods are combined to a system for pedestrian mobility analysis which comprises three layers. The *Sensor Layer* (1) monitors geo-coded sensor recordings on people's presence

and hands this episodic movement data in as input to the next layer. By use of standardized Open Geographic Consortium (OGC) compliant interfaces for data collection, we support seamless integration of various sensor technologies depending on the application requirements. The *Query Layer* (2) interacts with the user, who could ask for analyses within a given region and a certain time interval. Results are returned to the user in OGC conform Geography Markup Language (GML) format. The user query triggers the (3) *Analysis Layer* which utilizes the mobility model for pedestrian volume estimation.

The proposed approach is promising for location performance evaluation and attractor identification. Thus, it was successfully applied to numerous industrial applications: Zurich central train station, the zoo of Duisburg (Germany) and a football stadium (Stade des Costières Nîmes, France).

Acknowledgements

It would not have been possible to write the thesis at-hand without the support of the kind people around me.

This thesis would not have been possible without the support and advices from my principal supervisor Stefan Wrobel and my colleagues Zhao Xu, Michael May Gennady and Natalia Andrienko and Kristian Kersting. Without the support of Dirk Hecker during the last months, I would not have finished my project in the desired time.

The research presented in this thesis was project-driven and thus based on application scenarios. Therefore real-world data on people's presence had to be gathered. This is an undesirable task under cold, rainy or snowy weather conditions, often performed by myself as well. Foremost as this step was performed manually in the beginning and for this I would like to kindly thank all the students and colleagues from Mobility Mining and other groups from Knowledge Discovery for their hard work.

Also thanks to our project partner Swiss Poster Research Plus. I thank my colleague Terence Dörflinger for allowing me to gather first experience in project management. Furthermore I thank him for his support with computer hardware, my first computers were inherited by him. Also Jörg Kindermann did a great job in providing all the cluster capacities, required in various industrial applications. Karl-Heinz Sylla helped me with versioning of source files and documents and recommended literature on software development.

In a discussion with Thomas Kubitza from University Duisburg Essen I was convinced to use Bluetooth scanners for pedestrian monitoring. During the phase of assembling and first trials I received initial support at our institute by Reiner Frings, Robert Hofmann, Jochen Winzer, Günter Reuter and Dirk Hecker. Thank you for supporting me in application of this emerging technology, despite any skepticism whether the sensors are representative enough or at all industrially applicable.

Many fruitful discussions with Nico Van de Weghe, Mathias Versichele and Tijs Neutens who conduct the Bluetooth tracking experiments at University Ghent inspired most of the analysis performed in Chapter 4 of this thesis. Together with Reiner Frings, Hermann Streich, Timothy Ellersiek and Iulian Peca we installed this tracking technology at many industrial pedestrian monitoring scenarios and the experiences helped improving the hardware and raised interesting questions. The data collected at three of these numerous scenarios had impact on this thesis. Namely these were the data collection at the zoological garden in Duisburg (Germany), the data collection at the multi-purpose arena in Düsseldorf and finally the collection at Stade des Costères Nîmes (France) which was supported by our partner verint (in the European project ESS consortium). Thanks to all involved project partners.

The discussion and exchange in the European *Emergency Support System* (ESS) project was very fruitful. Especially I want to thank the project partners who implemented the ESS user interface to the hereby presented system for pedestrian mobility analysis. Also conference calls with our project partner intergraph (Czech) led to practically applicable software interfaces of our system, described in Chapter 5.

The European Comission project ESS (Emergency Support System) played an important role for this thesis, therefore I thank to the involved colleagues Haolin Zhi, Iulian Peca, Christian Pölitz, Natalia and Gennady Andrienko and Hans Voss.

During my work I had the chance to express my ideas to the European research community on spatial data mining at meetings of the GeoPKDD, MODAP and MOVE network. For these discussions I particularly want to thank to Natalia and Gennady Andrienko, Maria Damiani, Nico van De Weghe, Stefan Van der Speek, Yannis Theodoridis, Ralf Hartmut Güting, Monica Wachowitz, Daniel Orelana, Nikos Pelekis, Yücel Saygin, Fosca Gianotti, Dino Pedreschi, Anna Monreale, Dirk Helbing, Tijs Neutens, Mathias Versichele, Christine Kopp, Michael May, to name just a few.

I also want to thank the pedestrian simulation group at Jülich, foremost Armin Seyfried, Ulrich Kemloh and Bernhard Steffen for intensive discussions on the physical view on pedestrian movement, and agent based models. I also had a encouraging discussion with Dirk Helbing.

After hard efforts of persuasion, I created strong internal confidence to the Bluetooth tracking technology. I am glad to see my colleagues pushing the topic further, and stimulating usage of the hereby developed sensor technology in current research projects. I therefore thank them for the appreciation of my work, which for sure increases the impact of this thesis. The integration of the technology in a flexible, easily deployable, proactive, real-time pedestrian monitoring system has been my incentive for the last years to overcome the manual counting which I toiled for in the beginning of my time as a PhD candidate.

I thank to the students who I supervised during my thesis and who stimulated my work. Meena Shehata for his support on the earliest pedestrian simulation implementations (the F.A.S.T. algorithm of Tobias Kretz), and Haolin Zhi for the work on the interfaces of the ESS system, Patrick Utsch for carying out the experiments on Bluetooth fingerprinting and Simona Florescu for parts of the Bayesian network implementation. Moreover, Timothy Ellersiek for the statistical analysis and data acquisition of the zoo dataset.

I thank the "kdml doktoranden"-group, which assembles at a weekly base and gave critical feedback on status reports and presentation rehearsals, namely: Shankar Vembu, Frank Reichartz, Mario Boley, Christine Kopp, Sebastian Bothe, Anja Pilz, Marion Neumann, Dennis Wegener, Christian Pölitz, Babak Ahmadi, Mirwaes Wahabzadi, Hannes Korte, Fabian Hadiji and Daniel Paurat. Further fruitful discussions took place with Thomas Gärtner, Axel Poigné, Tamás Horvath, Barbara Krausz and I would like to thank them for that. At the Dagstuhl seminar no. 12331 in 2012 I head inspiring talks with Marco Gruteser and Michael Marhöfer, whom I like to thank.

I thank to Stefan Wrobel and Gennady and Natalia Andrienko for giving me the opportunity to work at the University of Bonn in parallel to the work at Fraunhofer. In my sadly too few visits at the Uni Bonn I had interesting talks to Daniel Seidel, Christian Pölitz, Jens Behley and Volker Steinhage who I kindly thank.

Thanks also to my colleagues from the Mobility Mining group Michael Mock, Dirk Hecker, Eike Stuckert, Hans-Hermann Streich, Hendrik Stange, Robert Spindler, Christine Kopp, Daniel Schulz, Detlef Geppert, Jutta Maas-Dittmann and Sebas-

tian Bothe for their support and discussion.

I thank Jutta Maas-Dittmann, Robert Spindler and Christine Kopp for being my room mates. It's not always easy, since I like warm temperatures in the office. Ahmet Ocakli, for respecting my preference for `vi`, though being on the emacs side of the LINUX world. In the technical discussions with you I felt as a computer scientist.

I thank my friend and former teacher Gerhard (Willi) Weber for raising my interest in operation research. The traffic quantity estimation task, which is subject of this thesis, has been an operation research problem. The hereby presented methods utilize methods of machine learning and spatial data mining. Nevertheless, it is important for a scientific dialog, to communicate the results and approaches interdisciplinary.

For their willingness and patience to proof-read and return feedback on my thesis I thank Simona Florescu, Zhao Xu, Georg Fuchs, Christine Kopp, Michael May and Dirk Hecker.

Lastly I would like to express my thanks to Simona for her personal support and her great patience at all times.

Contents

List of Figures

Chapter 1
Introduction

"We are happy to observe an increasing frequency of these pedestrian tours: to walk, is, beyond all comparison, the most independent and advantageous mode of travelling "

—Robin Jarvis[1]

Contents

Since the last decades, the proliferation of feature-rich and inexpensive computer hardware as well as its continuous evolution had a significant impact on the use and development of optimization software, logistical analysis and operations research in general [Maros & Khaliq 2002]. Specifically, processing power approximately doubles every 12 months (Moore's law [Moore 1965]), as does data storage capability every 18 months (Parkinson's law [Parkinson 1957]). This development came along with the evolution of highly sophisticated sensor technologies (e.g. motes). The tracking of moving objects and the foundation of a new research area, namely the *Spatio-Temporal Data Mining*, was the consequence. This thesis focuses on the primal *Operations Research* problem of modelling people's mobility with *Spatial Data Mining* methods. Therefore, this work tackles the estimation of pedestrian quantities at unobserved locations. This comprises (1) contribution of novel algorithms for pedestrian quantity estimation which for the first time incorporate knowledge on pedestrian movement patterns, (2) improvement of data acquisition and movement pattern analysis, (3) software integration and (4) extensive real-world applications. This chapter highlights the research questions, gives a brief introduction to challenges as well as to the state-of-the-art. Afterwards we express the author's contribution and an introduction to the application domain.

[1] British author, born 1963, Romantic Writing and Pedestrian Travel [Jarvis 1999]

1.1 Motivation and Research Questions

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 pedestrian 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 public restrooms depend on the quantity of persons, path-widths 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.

This thesis focuses on the estimation of traffic volumes for pedestrians within closed environments. Closed environments 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 which connect them with their surroundings. Prominent examples are train stations, terminals, shops, shopping malls, parks, as well as zoological gardens. Only in closed environments pedestrian quantity estimation can be analysed without any unexpected influences (for example the locations with arbitrary stay times from urban environments such as living houses and points of interest).

As shown in the previous examples, knowledge of pedestrian movement provides indispensable benefits to safety, marketing, as well as to infrastructural applications. Therefore, over the past years, many sensor technologies have been developed to fetch empirical measurements and record pedestrian volumes (most popular ones are video surveillance, laser beams and Bluetooth sensors). However, empirical measurements are usually rare due to constraints, e.g., budget limitations. This raises the following research questions:

- How to track pedestrians in mixed indoor/outdoor environments?
- How can values on pedestrian quantities be estimated from few empirical measurements?
- At which places should a constrained number of quantity sensors be located?
- How can the developed methods be used in practice?

Often, available information 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 movement patterns, although pedestrians generally show some move preferences, especially in closed environments, e.g., train stations. For example, consider a commuter on his daily path to work. Starting at home he uses the public train to reach his place of

work. In the train station it is more likely for the commuter to walk towards his desired platform than to walk to another one. He therefore chooses some routes more likely than others. Whereas predicting this individual behaviour is an interesting research field, this thesis focuses on the analysis and modelling of the total amount of pedestrians passing locations which is one aspect of pedestrian flow (besides others as density, route choice, flow mixing, etc.). Models of pedestrian flow can find application in problems related to *location evaluation*, *attractor identification* and *abnormality detection*. In these scenarios, the number of passing people (the so-called quantity) is a required input commonly used to evaluate locations [Liebig 2011].

Driven by the lack of (easily accessible) detailed floor plans for many buildings we are going to build a pedestrian quantity model that bases on the topology. Further inputs to the model are frequency counts and knowledge on movement patterns. Thus, the requirements for the application of quantity estimation method are:

- exclusively based on floor plan sketches and public available data sources,
- adjustable by quantity measurements,
- incorporating knowledge on movement pattern.

This thesis is driven by real-world scenarios and industrial projects. Hence, besides contributing two novel algorithms for pedestrian quantity estimation which enable movement pattern incorporation for the first time, this thesis also contributes novel methods for analysis of episodic movement recordings as well as an integrative software system.

This chapter proceeds with the challenges and the author's contributions, a brief description of the application scenarios, the outline of the thesis as well as a list of the author's publications.

1.2 Challenges

Many existing approaches deal with the research questions mentioned above i.e. pedestrian tracking, quantity estimation and sensor placement for traffic monitoring.

Regarding the first question for robust pedestrian monitoring many technologies evolved recently. Besides intrusive ones as mobile tracking applications (e.g. for smartphones) [Hoh *et al.* 2012, Florescu *et al.* 2012], these monitoring methods comprise non-intrusive ones as well: video surveillance [Masoud *et al.* 2001, Bertozzi *et al.* 2004], 3D laser scans [Schulz *et al.* 2003, Kräußling *et al.* 2008, Teichman & Thrun 2012] or analysis of mobile network performance data (which causes strong privacy objections [Giannotti & Pedreschi 2008]). The hardware for 3D laser scanner tracking is still expensive and thus not yet widely applied. The major drawbacks of video surveillance are the dependency on surrounding light, low image fidelity for far objects and occlusions by other objects. However, the approach of [Bruno & Delmastro 2003, Alt *et al.* 2009] is an appropriate way of solving the pedestrian monitoring task with recently evolved Bluetooth tracking technology. The wireless communication standard Bluetooth (used for example for linking intercoms with mobile phones)

is used for logging the occurrence of a Bluetooth enabled device close to a fixed stationary sensor. At a sensor location the aggregated numbers of traversing Bluetooth enabled devices is monitored. Usage of multiple sensors allows reconstruction of their individual transition times and movement patterns. However formal concepts and methods for analysis of the recorded data are still rare. Furthermore, the spatio-temporal representativeness is still not deeply studied.

The estimation of pedestrian quantities, i.e. the second research question, has been mainly discussed in the literature for vehicular traffic modelling. Among the existing approaches (Geographical Weighted Regression [Zhao & Park 2004], k-Nearest-Neighbour [Gong & Wang 2002], Gaussian Maximum Likelihood [Lam *et al.* 2006] and Kriging [Wang & Kockelmann 2009]) the Spatial k-Nearest Neighbour method [May *et al.* 2008] provides best performance. However these methods lack integration of complete domain knowledge (i.e. expert knowledges on movement behaviour). Furthermore the algorithms are designed for vehicular, and thus large scale, traffic networks. The focus on pedestrian traffic and comparable small environments reveals erroneous estimations of the algorithms more easily. In contrast to cars on a street no pedestrians can disappear within a corridor of a train station. Moreover the main challenge posed by the task is to provide methods that comply the requirements (1) exclusively based on floor plan sketches and public available data sources, (2) adjustable by quantity measurements, (3) incorporating knowledge on movement pattern.

The third question for sensor placement received recent attention in [Yang & Zhou 1998, Mínguez *et al.* 2010, Ng 2012, Gentili & Mirchandani 2012] and solves the task of placing sensors on traffic networks with the primal goal of estimating origin-destination matrices. As their goal is to estimate the relation among traffic sources and sinks and not the traffic flow among the traffic network, the objective is different and a novel method for sensor placement is required.

In detail, the contributions are described in the next section as well as in Chapters 3 and 4. In Chapter 2 we set the fundamentals for the hereby presented concepts. Application scenarios are briefly described in Section 1.3 and the very details of the software integration can be found in Chapter 5, followed by the project-driven research questions and analysis which are the focus of Chapter 6.

1.3 Contributions

Throughout this thesis we show that the proposed methods for pedestrian quantity estimation incorporating movement patterns overcome the weaknesses of the other related works. Also, our approach successfully meets the challenges (discussed in Section 1.2). We validate our approach against related state-of-the-art methods. With both synthetic ground truth as well as real-world data, the presented methods obtain higher accuracy (measured with mean absolute error). We furthermore enhance state-of-the-art person tracking technology (Bluetooth tracking). Furthermore, we develop visual analytics methods for analyzing *Episodic Movement Data* in terms of spatial correlations and temporal similarities.

We incorporate these functionalities in a software system for real-world applications which is successfully applied in industrial application scenarios. Briefly, the

contributions made in this thesis are:

- Novel methods for pedestrian quantity estimation which incorporate not just traffic network and sparse sensor readings, but allow incorporation of expert knowledge represented by movement patterns or heuristics on movement patterns.

- Analysis methods for a novel mobility data type, *Episodic Movement Data* which becomes prominent, e.g. by proliferation of Bluetooth tracking, RFID tracking, location based social networks or billing records. This comprises (1) analysis of spatial correlations and (2) temporal similarities.

- Integration of presented methods in a software framework adjusted to the requirements from real-world application scenarios. Application of the software system to real-world scenarios. Deployment of this software system within an European emergency support system.

We disseminated our contributions in the *data mining* and *machine learning* research field as well as in the *operations research* field by publishing our research in major research volumes (see Section 1.6).

1.4 Applications

The contributions of this thesis are driven by industrial applications and real-world scenarios. Thus, the requirements to the quantity estimation methods evolved from the applications. Additionally, existing acquisition and analysis methods have to be adopted and novel ones coping with the application requirements have to be found. The use cases we focused in these projects are *Location Evaluation*, *Attractor Identification* and *Abnormality Detection*. The application context is depicted in Figure 1.1. This

Figure 1.1: Three Pedestrian Quantity Estimation Application Scenarios: Location Evaluation for indoor poster advertisement (image credits: Hendrik Stange), Attractor Identification at the zoo (Duisburg) (image credits: nexuna@Flickr), Abnormality Detection in a soccer stadium (Nîmes).

section gives an overview of the different use cases and the related scenarios. We provide more details in Chapter 6.

1.4.1 Location Evaluation

In the first scenario, the *Billboard Location Evaluation Scenario*, an analyst is in need of pedestrian quantities in order to model the "visit potential" [Körner *et al.* 2010]. Thus, the pedestrian quantities are computed based on sensor readings by an expert and the resulting estimations are handed out to a domain expert in order to estimate visit potential based on the previous model. Thus, the pedestrian quantity estimation system has one expert user which incorporates input data given by *topological data* and *sensor readings*. Furthermore, since the expert performs this task, he achieved domain expert knowledge which is included in his estimations. The expert is a domain expert, and therefore knows how to work with geographical information systems. He utilizes software tools in order to estimate the quantities at unmeasured locations. The resulting model is visualised by the expert and in the eventual case of in-plausibility the input data or the software parameters will be adjusted. The resulting model can be stored in a generic transferable format, which is not just understandable by this expert user.

1.4.2 Attractor Identification

In the *Visitor Monitoring Scenario*, the visitors of a zoo are subject to analysis. For the zoo, the knowledge on the attractiveness of the compounds and shops is very interesting. This can be measured either in the quantity of persons or their stay times at various locations in the zoo. In order to evaluate and plan the signage of the zoo it is also important to analyse the dependencies within the visitor trajectories. However, a system for pedestrian quantity estimation is just a small part of the whole mobility analysis required for visitor monitoring in a zoo. The analysis needs to be possible without expert knowledge on geographic information systems.

1.4.3 Abnormality Detection

The *Event Monitoring Scenario* is integrated in an Emergency Support System. This Emergency Support System (ESS) [ESS 2010] is a software framework (developed by a European consortium) that supports forces in crisis management. Whenever an incident happens, the system is rapidly deployable and provides a comprehensive information system at the command post. In order to deal with different incidents, the system uses various technologies for data exchange. Data is collected from heterogeneous sensors, including video surveillance, chemical probes (gas, water, temperature) as well as geographical positions of the forces (fire fighters and police men) as well as their vehicles. All sensor readings are integrated in the system. The experts at the command post (the technical commanders) operate the system and depict the readings on a map. Furthermore, the locations of the forces are also shown on the map. Therefore the Emergency Support System enables the decision makers to base its decision on various incoming data. To assist this process multiple expert systems are included in the ESS software via standardized interfaces. These are not just focusing on people's mobility but also on prediction of gas (as well as fire or water) proliferation. We contribute to the system with pedestrian quantity estimation included in

a historical data analysis module. Thus, based on previously recorded (normal) data we return spatio-temporal pedestrian quantities to the expert who may use it in order to understand how people tend to use the provided infrastructure. The analysis of the recorded data is shown in the applications (Chapter 6) and the requirements posed by this use case and how to cope with them are explained in Chapter 5.

1.5 Outline

The remainder of the thesis at hand is structured into six proceeding chapters as follows.

- **Chapter 2: Pedestrian Mobility Fundamentals**
  This chapter discusses the specifics of pedestrian movement, its digital representation and models. We describe how motivated individual mobility leads to patterns in mobility recordings and present models for their representation. Furthermore, various pedestrian models will be discussed which describe different aspects of mobility, based on preliminary assumptions.

- **Chapter 3: Pedestrian Quantity Estimation Using Movement Patterns**
  Utilizing the preliminary work, presented in Chapter 2, this chapter focuses on pedestrian quantity estimation using movement patterns. We present the two contributed complementary regression approaches i.e. Least Squares Regression (LSR) and Gaussian Process Regression (GPR) to tackle the task with different input data and constraints on movement. We present the validation of the two contributed approaches as well.

- **Chapter 4: Movement Pattern Analysis based on Bluetooth Tracking Data**
  In this chapter, we describe how to analyse the correlations (caused by the movement patterns) in the observation data. Thus, we introduce a recently evolved tracking technology which records *Episodic Movement Data* on mobility and describe methods for analysis of this sparse data.

- **Chapter 5: A System for Pedestrian Mobility Analysis**
  Utilizing the previously presented work, this chapter focuses on the design of a *Pedestrian Monitoring System*. According to the requirement analysis we contribute a comprehensive approach: based on *Episodic Movement Data* we illustrate recording technologies for pedestrian mobility. These sensor measurements become input for data analysis that estimates traffic flow for unobserved locations using regression models. All components, i.e. (1) user interaction, (2) empirical data recording and (3) data analysis, are combined to our system for pedestrian mobility analysis. Whereas this section focuses on theoretical aspects of the system and validates the presented algorithms, the following chapter describes successful industrial applications.

- **Chapter 6: Real World Application Scenarios**
  This chapter describes the application of the algorithms and methods introduced in previous chapters to the real world problems. These comprise *Location Evaluation*, *Attractor Identification* and *Abnormality Detection*. In detail, the real world

applications we apply our methods to, are the following real world scenarios: *Billboard Location Evaluation* for Swiss train stations, *Visitor Monitoring* at the zoo of Duisburg and *Event Monitoring* during a soccer event at Stade des Costières, Nîmes (France). Posed questions trigger our method choices. Thus, besides pedestrian quantity estimation, in this chapter we provide a comprehensive analysis of the recorded data sets.

- **Chapter 7: Summary** In this chapter we summarize our contribution and provide future work suggestions. We summarize the two approaches (LSR and GPR) and the real-world applications. We present the benefits and the possible improvements (and extensions) of our approach.

1.6 Publications

The contributions of this thesis have already been published in the following journal, conference and workshop publications. All papers contain a significant contribution from the author of this thesis.

- T. Liebig, Z. Xu, M. May and S. Wrobel. *Pedestrian Quantity Estimation with Trajectory Patterns*. In Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases ECML PKDD 2012, Part II, LNCS 7524, pages 629–643. Springer, 2012

- N. Andrienko, G. Andrienko, H. Stange, T. Liebig and D. Hecker. *Visual Analytics for Understanding Spatial Situations from Episodic Movement Data*. KI - Künstliche Intelligenz, pages 241–251, 2012.

- T. Liebig and Z. Xu. *Pedestrian monitoring system for indoor billboard evaluation*. Journal of Applied Operational Research, vol. 4, pages 28–36, 2012.

- T. Liebig, G. Andrienko and N. Andrienko. *Methods of Analysis of Episodic Movement Data*. In Mobile Tartu, pages 24–25, 2012

- T. Liebig, Z. Xu and M. May. *Incorporating Mobility Patterns in Pedestrian Quantity Estimation and Sensor Placement*. In J. Nin and D. Villatoro, editors, Proceedings of the First International Workshop on Citizen Sensor Networks CitiSens 2012, LNAI 7685, pages 67–80. Springer, 2013

- T. Liebig. *Trajectory Regression Model for Indoor Pedestrian Flow Analysis on Billboard Evaluation*. In Proc. of the Third International Conference on Applied Operation Research - ICAOR'11, pages 289–300. Tadbir Operational Research Group Ltd., 2011.

- T. Liebig, H. Stange, D. Hecker, M. May, C. Körner and U. Hofmann. *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.

- T. Liebig, C. Körner and M. May. *Fast Visual Trajectory Analysis Using Spatial Bayesian Networks*. In ICDM Workshops, pages 668–673. IEEE Computer Society, 2009.

Supplementary contributions appeared in the following conference and workshop publications.

- T. Liebig and A. U. Kemloh Wagoum. *Modelling Microscopic Pedestrian Mobility Using Bluetooth*. In Proc. of the Fourth International Conference on Agents and Artificial Intelligience - ICAART'12, pages 270–275. SciTePress, 2012
- T. Ellersiek, T. Liebig, D. Hecker and C. Körner. *Analyse von raum-zeitlichen Bewegungsmustern auf Basis von Bluetooth-Sensoren*. In Angewandte Geoinformatik 2012 - Beiträge zum 24. AGIT-Symposium Salzburg, pages 260–269, Berlin, 2012. Wichmann.
- P. Utsch and T. Liebig. *Monitoring Microscopic Pedestrian Mobility Using Bluetooth*. In Proceedings of the 8th International Conference on Intelligient Environments, pages 173–177. IEEE Press, 2012
- H. Stange, T. Liebig, D. Hecker, G. Andrienko and N. Andrienko. *Analytical Workflow of Monitoring Human Mobility in Big Event Settings using Bluetooth*. In ISA 2011, pages 51–58. ACM, 2011.
- T. Liebig, C. Körner and M. May. *Scalable Sparse Bayesian Network Learning for Spatial Applications*. In ICDM Workshops, pages 420–425. IEEE Computer Society, 2008.

Chapter 2

Pedestrian Mobility Fundamentals

"While the individual man is an insoluble puzzle, in the aggregate he becomes a mathematical certainty. You can, for example, never foretell what any one man will be up to, but you can say with precision what an average number will be up to. Individuals vary, but percentages remain constant. So says the statistician."

—Sir Arthur Conan Doyle[1]

Contents

[1] Scottish physician and writer, 1859–1930, The Sign of Four [Doyle 1890]

This chapter discusses the specifics of pedestrian movement, its digital representation and models. We provide an introduction to spatio-temporal geography and its digital data representation. Furthermore, we describe how motivated individual mobility leads to patterns in mobility recordings and present models for their representation. Furthermore, various pedestrian models will be discussed which describe different aspects of mobility based on preliminary assumptions. In next chapters we present how to monitor and analyse these patterns and present, validate and apply novel approaches that incorporate them for quantity estimation.

2.1 Empirical Facts on Pedestrian Mobility

Pedestrians are not homogeneous. Every single person is characterized by its sex, its age, its height and space requirements as well as its biomechanics and its physiology [Weidmann 1993]. As an example we reflect briefly the distribution of walking speed. A literature overview and discussion of existing empirical studies concerning facts of pedestrian movement is given in [Weidmann 1993]. Thus, the average pedestrian speed is denoted by $1.34m/s$ ($4.83km/h$) which depends on seven influencing factors:

- **sex:** Male Pedestrian are a bit faster with $1.41m/s$ than female ones $1.27m/s$.
- **age:** Highest speed is reached with age of 20 years. After reaching 50 years walking speed gets significantly lower.
- **purpose:** Workers walk fastest with $1.61m/s$, commuters are a bit slower with $1.49m/s$. When shopping the pedestrian walk with $1.16m/s$ and during leisure time they are slowest with $1.10m/s$.
- **time:** In the morning the speed is highest.
- **temperature:** The temperature has also impact on pedestrian speed. Whereas the speed is reduced to 92% of the average at 25 degree Celsius, it reaches 109% at 0 degree (average speeds measured at 15 degree Celsius [Weidmann 1993]).
- **pavement length:** No impact of the pavement length was observed in the plane but on stairways.
- **density:** The more densely a corridor is crowded, the lower average speeds become.

Whilst these factors have high impact on individual movement, the pedestrian decisions and their behaviour also influence the individual movement. These influences are subject to the *hierarchy of motion* [Hoogendoorn *et al.* 2002]. This hierarchy consists of three routing levels representing different levels of path planning, see Figure 2.1. The three tiers have local interactions with each other, marked in the figure by the black arrows and described below.

- **Strategic Level:** In the strategic level the pedestrian chooses its target and the strategy how to get there. This is the self estimated best route, among a collection of different alternatives. This can be done based on experience. Examples could be the global shortest path or the familiar path to a given destination.
- **Tactical Level:** Short-terms decisions are made at the tactical level, avoiding jams or switching to a faster route for instance. Thus, the person chooses the path to avoid obstacles. Basic rules for motions are defined at the tactical level, these include accelerating, decelerating, stopping.
- **Operational Level:** In the operational level, the motion to the next intermediate point is performed e.g. decision for a movement direction and speed or planning of the next step.

To sum up, the planning process for pedestrian motion divides into *target selection*, *path selection* and *real-time replanning*. Thus, dependencies and correlations occur among the movement of every single person. An aggregation of all people's movement returns persistent movement habits, therefore mobility patterns arise. The next chapter introduces novel methods for traffic quantity estimation that make use of these mobility patterns, and in Chapter 6 we show applications of these methods.

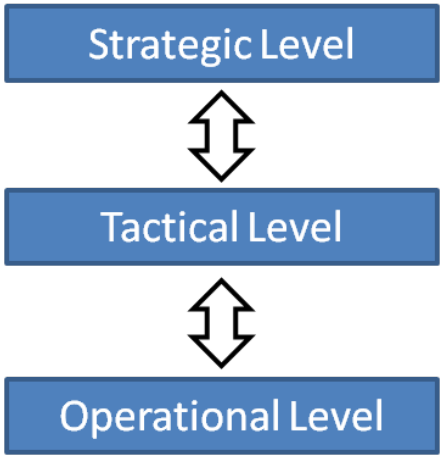

Figure 2.1: Hierarchy of motion [Hoogendoorn *et al.* 2002].

2.1.1 Target Selection

Pedestrian movement is not a random walk, but aims at specific goals. For example, consider a commuter on his daily path to work. Starting at home he uses the public train to reach his place of work. In the train station it is more likely for the commuter to walk towards his desired platform than to walk to another one. He therefore chooses some routes more likely than others. Whereas predicting this individual behaviour is an interesting research field, this thesis focuses on analysis and modelling of the total amount of pedestrians passing locations which is one aspect of pedestrian flow (besides others as density, route choice, flow mixing, etc.). In [Helbing 1997] the selection

of the movement target is discussed. In case the pedestrian has multiple goals, for an example during shopping, he needs to walk to one goal after the other. Usually, the selection and order of the targets aims at a minimization of the total path-length. The impact of flocking on target selection is presented in [Heliövaara *et al.* 2011]. There, a group of people is observed when leaving a room, the participants of the experiment having the choice between two doors. Surprisingly the decision is not just based on the path distance but also on the queue length. Furthermore, the persons walking in the front chose the further exit in order to give more space to the ones behind. The experiment was carried out with a survey and the participants told that an additional factor for their exit choice was the starting position within the corridor. All of these path choices are factors to be considered when tackling the estimation of pedestrian quantities [Hoogendoorn *et al.* 2002].

2.1.2 Path Selection

Besides target selection the route choice is important. The hierarchy of motion [Hoogendoorn *et al.* 2002] (for convenience depicted in Figure 2.1) describes that the route choice is performed in the tactical routing level, however [Helbing 1997] states that the route choice is dynamic and may be replanned in case of unexpected attractors (e.g. shop windows). These decisions are also influenced by the operational routing level [Hoogendoorn *et al.* 2002]. In literature, most experiments justify the assumption that most of the people prefer the quickest path [Borgers & Timmermans 1986, Guo & Huang 2011, Hoogendoorn *et al.* 2002]. The movement is planned with intermediate targets, claimed by [Helbing 1997, Hoogendoorn *et al.* 2002]. Also in [Helbing 1997] is claimed that the pedestrians prefer eye-catching locations as intermediate targets. However, in case of rain, or slippery ground, the pedestrians are more likely to accept detours [Helbing 1997]. Recent work [Kemloh Wagoum & Seyfried 2011] focusses on simulation of evacuation with dynamic route choice. In case of train stations a laser beam experiment proofed the shortest path assumption for this particular scenario [Li *et al.* 2008]. In their study about 2,500 traces of train station visitors were analysed. We utilize this justification to model pedestrian quantities of train station visitors in Chapter 6. However, for persons performing leisure activities, e.g., tourists or zoo visitors, these assumptions do not hold in general [Kretz 2009, Helbing 1997]. In Chapter 4 we describe how Bluetooth tracking (Section 2.3.4.2) captures movement behaviour even in these scenarios. And in Chapter 6 we present successful real-world application using movement patterns and Bluetooth tracking at a soccer stadium in Nîmes and the zoological garden of Duisburg.

2.1.3 Group Movement

Usually the distance between the pedestrians depends on the density of the people [Helbing 1997]. But often pedestrians walk in groups of two, three or four persons [Peters & Ennis 2009]. In these groups, the distance among the persons is lower than the average pedestrian distance (i.e., interpersonal space). The video surveillance study performed in [Peters & Ennis 2009] also analyses typical formations of the group. The results of this study reveal that people in a group do not walk shoulder

to shoulder but adopt a staggered formation. In result, the frontal aspect of the group can be dynamically changed without dropping group members from the others' interpersonal space. Even without any obstacles and other constraints, the distance among the group members does not remain constant. Thus, the shape of the group changes and complex movement intersections among multiple groups arise [Peters & Ennis 2009].

2.1.4 Collective Behaviour

In pathways with pedestrians walking in opposite directions, the different movement streams separate themselves. The higher the pedestrian density gets, the more these streams grow, till just two remain [Helbing 1997]. Thus, the loss of efficiency (loss of maximum traffic frequency) in a pedestrian environment due to two-way traffic is of about 4% (in case of equal frequencies). Often, narrow pathways are passed in alternating direction by throngs. The frequency for the oscillation of the movement direction increases by the width of the corridor and the length of the pathway [Helbing 1997].

Another collective behaviour is the creation of dirt tracks. This phenomenon is studied in [Helbing 1997]. The creation of dirt tracks results from the avoidance of pedestrians of uneven ground and the preference to walkable areas. These traffic systems possess characteristic properties [Helbing 1997]. In case three paths meet at a junction, the intersection is not rectlineal but the paths turn into each other smoothly. In contrast, intersections of four paths are often the crossing of two pathways and the rectlineal extent is preserved. Moreover, dirt tracks with a high traffic frequency are wider than less popular ones. In case a dirt track becomes un-walkable, a novel one emerges parallel to the existing one.

In conclusion, *target selection*, *path selection*, *group movement* and *collective behaviour* are all empirical facts on pedestrian movement, studied in the literature, which influence people's motion [Hoogendoorn *et al.* 2002]. Therefore the quantity estimation methods as well as its incorporated expert knowledge should also reflect these facts.

2.2 Spatio-Temporal Geography

Whereas previous section described the specifics of pedestrian movement, this section addresses the necessary aspects for analysis of pedestrian movement. Thus, spatio-temporal reference systems are defined. Afterwards, database management systems for storage of this data are introduced, focussing on different aspects of spatial data. Integration of the data in software systems and mapping tools is commonly done via *Open Geographic Consortium protocols* which is a communication standard for spatial components. Finally, *Episodic Movement Data* and methods for its analysis are presented. Next sections tackle the monitoring of pedestrian movement and the state-of-the-art pedestrian mobility models.

2.2.1 Time Geography

In [Minkowski 1908] the fundamental relation between space and time is formulated. Whereas space was before considered as a three-dimensional homogeneous and isotropic extent, defined by Newton in [Newton *et al.* 1803], Minkowski firstly introduces a four-dimensional extrapolation of the prior three-dimensional one. He combines time t with the previously introduced spatial extent $X = (x, y, z)$ (or $X = (x, y)$ for two dimensional coordinate systems) by introducing ict as additional dimension, with ($i^2 = -1$ and c is the maximal speed of light). Thus the sphere that a light-flash creates in space $x^2 + y^2 + z^2 = r^2$ becomes in four dimensions:

$$\begin{aligned} x^2 + y^2 + z^2 &= r^2 = c^2 t^2 \\ x^2 + y^2 + z^2 + (ict)^2 &= 0,\ i^2 = -1\ . \end{aligned} \tag{2.1}$$

In four dimensions, this equation describes a cone, depicted in Figure 2.2. Hence

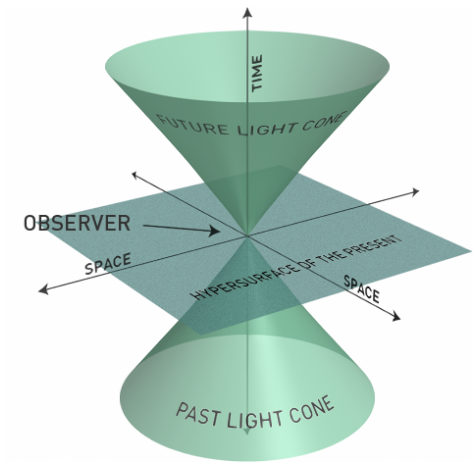

Figure 2.2: Light Cone in Minkowski Space-Time (image credits: [Wikimedia Commons]).

it is called the *light-cone*. The coordinates in this Minkowski space (x, y, z, ict) are called *Minkowski Coordinates* whereas (x, y, z, ct) are called *Galilei Coordinates* [Schmutzer 1989]. Having the reasonable assumption, that light-speed is the limit for all speeds, a point at (x, y, z, ict) may not be affected by anything except the lower half cone, therefore called *past*. The upper half cone describes its *future* [Schmutzer 1989], compare Figure 2.2. This close relation among space and time is also expressed in the definition of the spatial unit length *meter [m]*.

Definition 1 (meter) *A meter [m] is the distance light passes in 1/299.792.458 seconds in vacuum [National Institute of Standards and Technology 2008].*

Note that this definition implies that the speed of light c is given by $299.792.458 m/s$. Whilst this space still possesses the Euclidian geometry, which describes a rectlineal space, later the curvlineal Riemann geometry was considered which allows explanation of gravity effects.

However, for this work, the gravity effects and relativistic influences of mass objects are not relevant. Therefore the four-dimensional space-time with rectlineal Euclidian geometry is sufficient for pedestrian modelling. Whereas physicists apply this model since 1908, in 1970 the space-time was firstly incorporated in geography [Hägerstrand 1970] using the term *Time-Geography* to visualize and analyse the motion of pedestrians. Thus the spatial coordinate system, in geography often represented by x, y due to the two-dimensionality of maps, is extended by time t as a third dimension. This results in a continuous three-dimensional line-visualization of individual movement. In physics it is common to call this line a *trajectory* [Schmutzer 1989]. Introducing a maximal speed for the motion of people (similarly to the maximum speed of light above) the transition possibilities among two points in (x_1, y_1, t_1) and (x_2, y_2, t_2) results in a so called *space-time prism* [Lenntorp 1976]. This is a volume in space-time which is formed by the intersection of the *future* of point (x_1, y_1, t_1) and the *past* of (x_2, y_2, t_2), compare Figure 2.3. When projecting this prism to the x, y plane it defines the *possible path area*.

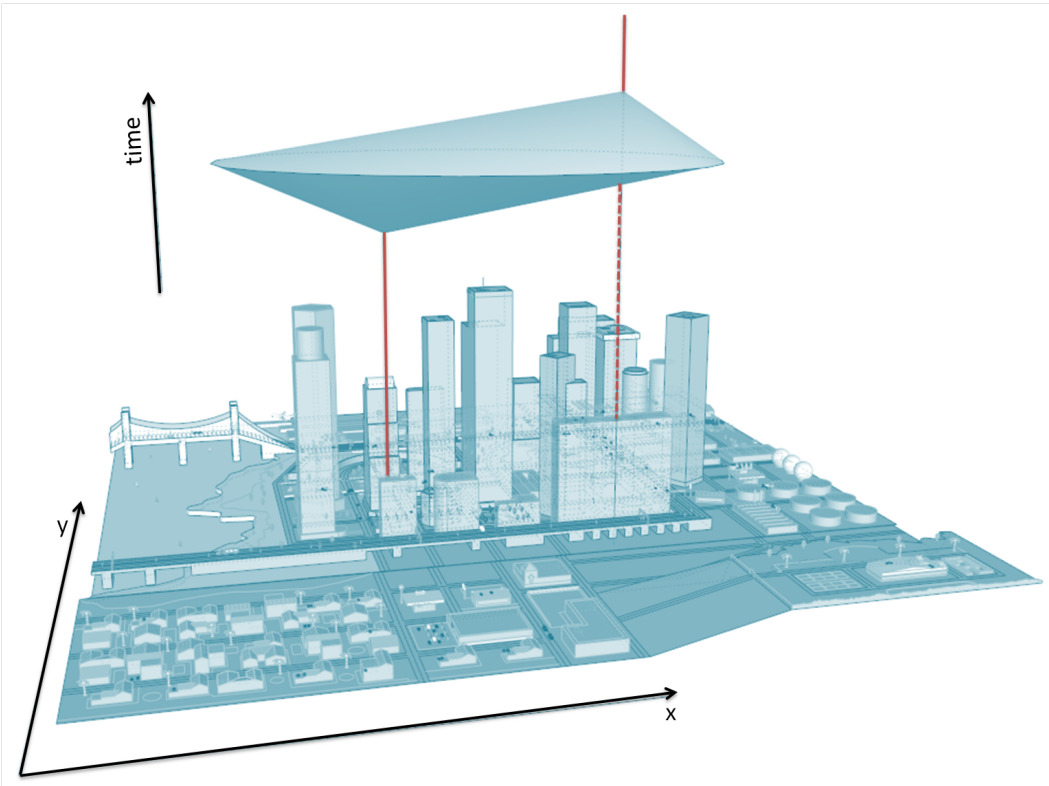

Figure 2.3: Example of space-time prism representation of movement among two places in space-time. The red lines mark that the person spends some time at the left building and later on stays in the right building. The transition in between the two spatio-temporal locations is bounded by the space-time prism in the upper part of the picture.

A trajectory of a moving point, which represents its positions in space-time, can then be defined as:

Definition 2 (Trajectory) *A set of space-time points (x, y, z, t) is called trajectory S (or world-line) of a moving object, if for every contained time-stamp t, exactly one spatial point (x, y, z) is contained in S (uniqueness) and temporally subsequent points are contained in their light-cones (continuity). [Minkowski 1909]*

Thus, in a trajectory the spatial component X can be regarded as a function of time $X = f(t)$. In contrast to a moving point, the trajectory of a mass object is constrained by physical properties, e.g. *inertia*, *impulse*, *spin* and *gravity*. This results in stronger constraints for the continuity of trajectories. It can be expected to be *smooth*, i.e., $f(t)$ is continuously differentiable $\partial f(t)/\partial t$. Recent work on kinetic space-time prisms [Kuijpers *et al.* 2011] incorporates the physical behaviour of mass objects.

Time is a continuous and linear extent with unit of a second [s]. Possible spatial reference systems for the spatial component of the space-time points are so-called geo-reference systems, e.g. *WGS 84* [National Imagery and Mapping Agency 2000] which locates any point on earth by its ellipsoidical coordinate triple longitude, latitude and altitude. As WGS 84 became part of Open Geographic Consortium specifications, this geo-reference system is widely used. Another often used Cartesian coordinate system is given by the map itself, having the x and y axis perpendicular in the plane of the map image (and eventually the third z axis pointing orthogonal from the map surface).

2.2.2 Digital Data Representation

Automatic processing of trajectories and analysis of movement requires digital data storage. Two possibilities for digital storage of spatial data evolved: (1) grid and (2) vector representation. Grid representation aggregates spaces and straight contours are approximated by tessellations (see Figure 2.4), whereas vector representation preserves the spatial contours (see Figure 2.5). Vectorized data needs less memory and it allows easy integration of additional dimensions. In *raster space* this would imply transition from pixel to voxel-space. Furthermore, vector models provide easy mapping of data between various geographic coordinate systems with different range or precision. Primitive spatial vector object types are *point*, *line*, *area*, *network*, *text* and their *compound* [Bartelme 1995] shown in Figure 2.5. For the exchange and storage of spatio-temporal data the Open Geographic Consortium defined open file format standards, protocols and interfaces. Most popular ones are the Keyhole Markup Language (KML) which is an XML notation for description of spatial vector objects in WGS 84 [National Imagery and Mapping Agency 2000], or Geographic Markup Language (GML) which is mostly similar to KML but can describe arbitrary dimensions. The interfaces and their associated protocols define the connections among software modules or devices. For example, the transmission of sensor readings that are located at a specific location should adhere the *Sensor Observation Service Protocol* (SOS), whereas the transmission of map information from a server to a map-service could be either done in vector format (using the Web Feature Service protocol) or in raster format (following the Web Map Service protocol). The creation of these open communication standards led to a modularization of previously proprietary (closed) geographical information systems.

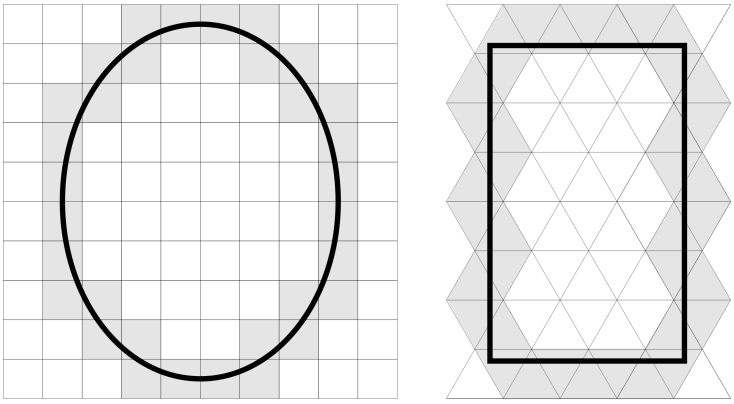

Figure 2.4: Tessellated Spatial Objects (image credits [Liebig 2007]).

2.2.3 Spatial Database Management Systems

These spatial primitives can be stored digitally in relational database management systems (e.g., using the PostgreSQL spatial extension or the proprietary Oracle Spatial data types), where every primitive is associated with a nested table that stores its spatial extent. However, this representation does not reflect the temporal aspects of time

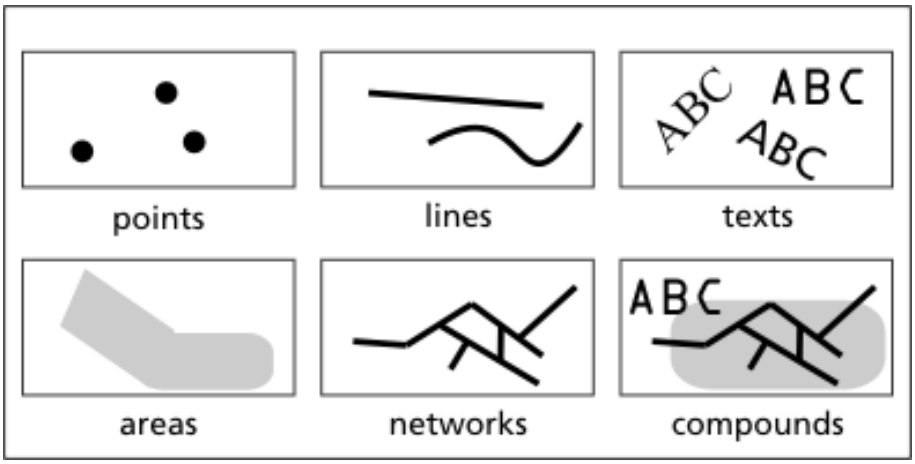

Figure 2.5: Spatial Vector Objects [Bartelme 1995] (image credits [Liebig 2007]).

geography. Thus, [Güting & Schneider 2005] introduced *moving object databases*. In this database system the conventional data types (integer, boolean, etc.) are extended by a time component. This describes the so-called *moving integer*, *moving boolean*, etc. Moreover, novel types are introduced, related to the vector geometry data types (e.g. *point* and *moving point*). While the conventional numeric or boolean types may change their value as a function of time in moving object databases, the attributes of spatial data (i.e. *size*, *shape*, *orientation* and *existence*) can also depend on time. The group of spatio-temporal operators like *intersection* or *unification* and these moving data types

are carefully defined to form an algebra on spatio-temporal objects. Therefore, all spatio-temporal computations can be described within the type system. A trajectory can be stored using a *moving point* data type. Further improvements of this approach were made by [Costa *et al.* 2007], who transformed this formalization into an extension for the functional programming language Haskell by creating TerraHs. This allows the seamless combination of the spatio-temporal algebra with algebra from other domains. Similar to the online analytical processing of database systems [Codd *et al.* 1993], Trajectory Data Warehouses [Marketos *et al.* 2008,Orlando *et al.* 2007] support an analyst to handle massive trajectory datasets. Therefore, trajectories are aggregated in the database according to the query and the results are subject for visual or analytical inspection [Leonardi *et al.* 2009].

2.2.4 Episodic Movement Data

The popularity of cellular phones and advances in information and sensor technologies lead the way towards new location recording techniques and thus new types of movement data. *Episodic Movement Data* [Andrienko *et al.* 2012] refers to data about spatial positions of moving objects where the time intervals between the measurements may be quite large and therefore the intermediate positions cannot be reliably reconstructed by means of interpolation, map matching, or other methods. Though trajectories are continuous in space-time, recordings often make some discretization either in space or time. Such *Episodic Movement Data* can also be called *temporally sparse*; however, this term is not very accurate since the temporal resolution of the data may greatly vary and occasionally be quite fine. There are multiple ways of data collection producing episodic movement data [Andrienko *et al.* 2012]:

- **Location based:** Positions of objects are recorded only when they come into the range of static sensors. The temporal resolution of the collected data depends on the coverage and density of the spatial distribution of the sensors. Possible recording technologies are cell based tracking methods, e.g., WLAN, GSM, Bluetooth. We refer in the next sections mostly to Bluetooth.
- **Activity based:** Positions of objects are recorded only at the times when they perform certain activities, for example, call by mobile phones, pay by credit cards or send posts to a community website.
- **Device based:** Positions are measured and recorded by mobile devices attached to the objects but this cannot be done sufficiently frequently, for example, due to the limited battery lives of the devices i.e. when tracking movements of wild animals.

Irrespective of the collection method we can identify three types of uncertainty (Figure 2.6). Firstly, the common type of uncertainty in any episodic movement data is the lack of information about the spatial positions of the objects between the recorded positions (continuity), which is caused by large time intervals between the recordings and by missed recordings. Secondly, a frequently occurring type of uncertainty is the imprecision of the recorded positions (accuracy). Thus, a sensor may detect an object within its range but may not be able to determine the exact coordinates of the object.

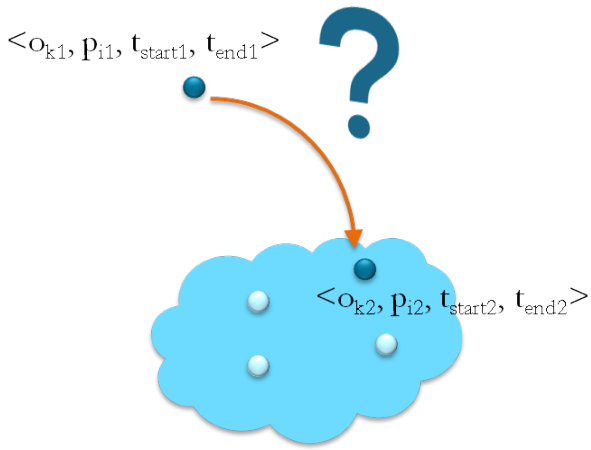

Figure 2.6: Three Common Uncertainties in Episodic Movement Data: orange arrow represents the uncertainty of the movement among two data points (continuity), blue cloud represents the uncertainty on (accuracy) and grey dots depict uncertainties on coverage.

For a mobile phone call, the localization precision may be the range of a certain antenna but not an exact point in space. Due to these uncertainties, episodic movement data cannot be represented as continuous trajectories, i.e., lines in the spatio-temporal continuum where known (measured) positions are linked by straight or curved segments. Third, the number of recorded objects (coverage) may also be uncertain due to the usage of a service or due to the utilized sensor technology. For example, one individual may carry two or more devices with Bluetooth transceivers, which will be registered by Bluetooth sensors as independent objects. On the other hand, the sensors only capture devices with activated Bluetooth services. The activation status may change while a device carrier moves from one sensor to another. Many of the existing visual and data mining methods designed for dealing with movement data are explicitly or implicitly based on the assumption of continuous object movement between the measured positions and are therefore not suitable for episodic data. Interpolation is obviously involved in visual representation of trajectories by continuous lines but it is also implicitly involved in computation of movement speeds, directions, and other attributes characterising the movement (these computations also assume that the positions are precise). The same holds for summarisation of movement data in the form of density or vector fields. Mining methods for finding patterns of relative or collective movement of two or more objects (e.g. meeting or flocking) also require fine-resolution data [Hai *et al.* 2012]. Since many of the existing methods are not applicable to episodic movement data, there is a need in finding suitable approaches to analysing this kind of data. Due to the uncertainties, episodic data are usually not suitable for studying the movement behaviours of individual objects. In order to overcome these shortcomings, we suggest (in [Andrienko *et al.* 2012, Stange *et al.* 2011, Liebig *et al.* 2012a, Liebig *et al.* 2013]) aggregation of many individual tracks to compensate for missing data and uncertainties in the spatial and temporal coverage.

By example of episodic movement data, this work motivates the utilisation of visual and computational methods to analyse complex data [Liebig *et al.* 2012a]. Visual analytics strives at multiplying the analytical power of both human and computer by finding effective ways to combine interactive visual techniques with algorithms for computational data analysis [Keim *et al.* 2008]. The key role of the visual techniques is to enable and promote human understanding of the data. Particularly, visual analytics can help in understanding the data for data mining tasks, such as distributions, features, clusters or patterns. Visual analytics approaches are applied to data and problems for which there are (yet) no purely automatic methods to deal with [Andrienko & Andrienko 2012]. By enabling human understanding, reasoning, and use of prior knowledge and experiences, visual analytics can help the analyst to find suitable ways for data analysis and problem solving, which, possibly, can later be fully or partly automated. Thus, visual analytics can drive the development and adaption of learning and mining algorithms. In the next section, we describe aggregation of episodic movement data and its visualization based on *visits* and *flows*.

2.2.4.1 Spatio-temporal aggregation

Episodic movement data consists of records including the following components: object identifier o_k, spatial position p_i, time t, and, possibly, other attributes. The spatial position may be specified directly by spatial (geographic) coordinates $p = (x, y)$ or $p = (x, y, z)$ or by referring to a sensor or location having a fixed position and dimension in space. A chronologically ordered sequence of positions of one moving object can be regarded as an abstract *trajectory* which is spatially and temporally discontinuous. For temporal aggregation of the data, time is divided into intervals. Depending on the application and analysis goals, the analyst may consider time as a line (i.e. linearly ordered set of moments) or as a cycle, e.g., daily, weekly, or yearly. Accordingly, the time intervals are defined on the line or within the chosen cycle. For spatial aggregation, it is necessary to define a finite set of places visited by the moving objects. Two different cases need to be distinguished:

- (1) The object positions in the data are limited to a finite set of predefined positions, such as positions of sensors or cells of a mobile phone network.
- (2) The object positions in the data are arbitrary. This is the case when the positions are received from mobile devices worn by the persons (i.e., mobile objects) and capable of measuring absolute spatial positions, such as GPS devices.

In the first case, the different positions from the data can be directly used as places for the aggregation. In the second case, spatial tessellation may give the required set of places (space compartments) for the aggregation. To successfully analyse the movement recordings at a higher spatial scale, the analyst may group neighbouring positions and define places as convex hulls or spatial buffers or Voronoi polygons [Voronoï 1908] around the groups. However, arbitrary divisions, such as regular grids, do not reflect the spatial distribution of the data. It is more appropriate to define space compartments so that they enclose existing spatial clusters of points. However, these clusters may have very different sizes and shapes, which has two disadvantages. First, it is computationally hard to automatically divide a territory into

arbitrarily shaped areas enclosing clusters. Second, the areas would differ much in their sizes, and the respective aggregates would be incomparable. Therefore, [Andrienko & Andrienko 2011] suggests a method that divides a territory into convex polygons of approximately equal sizes on the basis of point distribution. The method finds spatial clusters of points that can be enclosed by circles with a user-chosen radius. A concentration of points having a larger size and/or complex shape will be divided into several clusters. Next, a Voronoi tessellation [Dirichlet 1850, Voronoï 1908] is performed using these centroids of the clusters. The centroids are the points with the minimal average distance to the cluster members. They are usually located inside concentrations of points.

On the basis of a defined set of discrete places P, each trajectory T_o of a moving object o is represented by a sequence of visits $v_1, v_2, \ldots, v_n$ of places from P. We adopt the definition from [Körner *et al.* 2010] as follows:

Definition 3 (Visit) *A visit $v_i \in T_{o_k}$ is a tuple $< o_k, p_i, t_{start}, t_{end} >$, where o_k is the moving object, $p_i \in P$ is a discrete place, t_{start} is the starting time of the visit, and t_{end} is the ending time, with:*

- $t_{start} \leq t_{end}$,
- $\forall t \in [t_{start}, t_{end}] : T_{o_k}(t) \in p_i$,
- $T_{o_k}(t_{start} - \varepsilon) \notin p_i \wedge T_{o_k}(t_{end} + \varepsilon) \notin p_i$.

Complementary to this, each trajectory is also represented by a sequence of moves $m_1, m_2, \ldots, m_{n-1}$ where each move m_i is defined in this thesis as follows.

Definition 4 (Move) *A move m_i is a tuple $< o_k, p_i, p_{i+1}, t_0, t_{fin} >$ describing the transition of a moving object o_k from visit $v_i =< o_k, p_i, t_1, t_0 >$ to visit $v_{i+1} =< o_k, p_{i+1}, t_{fin}, t_2 >$, with:*

- $t_0 < t_{fin}$,
- $p_i \cap p_{i+1} = \emptyset$.

Here t_0 is the time moment when the move began (it equals t_{end} of the visit v_i of the place p_i) and t_{fin} is the time moment when the move finished (it equals t_{start} of the visit v_{i+1} of the place p_{i+1}). It should be borne in mind that consecutively visited places p_i and p_{i+1} in a discontinuous trajectory are not necessarily neighbours in space. Having a dual representation of trajectories, as sequences of *visits* and as sequences of *moves*, the data can be aggregated in two complementary ways [Andrienko *et al.* 2012]. First, for each place p_i and time interval Δt, the visits of this place from the interval are aggregated, i.e., the tuples $< o_k, p_i, t_{start}, t_{end} >$ where $\forall t : t_{start} \leq t \leq t_{end}$ and $t \in \Delta t$. The count of the visits and the count of different visitors (o_k) are computed. If the original data records include additional attributes, various statistics of these attributes can also be computed, such as minimum, maximum, average, median, etc. Hence, each place is characterized by two or more time series of aggregate values: *counts of visits*, *counts of visitors*, and, possibly, additional

statistics of the time intervals. The second way of aggregation is applied to *links*, i.e., pairs of places $< p_i, p_j >$ such that there is at least one move from p_i to p_j. For each link $< p_i, p_j >$ and time interval Δt, the moves from p_i to p_j from this interval are aggregated, i.e., the tuples $< o_k, p_i, p_j, t_0, t_{fin} >$ where $t_{fin} \in \Delta t$ (which means that only the moves that finished within the interval Δt are included). The count of the moves and the count of different objects that moved (o_k) are computed. If the original data include additional attributes, it is also possible to compute changes of the attribute values from t_0 to t_{fin} and then aggregate the changes by computing various statistics. Hence, each link is characterized by two or more time series of aggregate values: *counts of moves*, *counts of moving objects*, and, possibly, additional statistics of changes by the time intervals. Moreover, the counts can be defined for undirected movements NM_u, this comprises movement among locations p_i and p_j or vice versa. We are going to apply this *undirected counts of flows* in Chapter 3 and consecutive chapters of real-world applications. These two ways of aggregation support two classes of analysis tasks:

- Investigation of the presence of moving objects in different places and the temporal variation of the presence. The presence is expressed by the counts of visits and visitors in the places.
- Investigation of the flows (aggregate movements) of objects among different places and the temporal variation of the flows. The flows are represented by the counts of moves and moving objects for the links. These aggregate attributes are often referred to as flow magnitudes.

In both classes of tasks, the aggregated data can be viewed in two ways. Obviously, the data can be viewed as (1) time series associated with the places or links. The analyst can investigate the individual time series or groups of time series (e.g., clusters of similar time series) using existing methods for time series analysis. On the other hand, the data can be viewed as (2) a sequence of spatial situations associated with the time intervals.

A *spatial situation* is the distribution of the object presence or flows over the whole territory during a time interval. The different views on aggregated movement data are illustrated by maps in Figure 2.7. In Figure 2.7A and 2.7B, time series of aggregate values associated with two selected places (A) and with a selected link between two places (B) are represented by polygons where the horizontal dimension represents time and the height is proportional to the values in different time intervals. The places themselves are represented by ellipses and the link by a special symbol (further referred to as flow symbol) looking as a half of an arrow and pointing in the direction of the movement. Such half-arrow symbols are used to be able to represent flows between two places in two opposite directions. In Figure 2.7C and 2.7D, spatial situations in a selected time interval in terms of presence (C) and flows (D) are shown. The presence is shown by proportional heights of the bars drawn in the places and the flow magnitudes by proportional widths of the flow symbols. A map where aggregated movement is shown by flow symbols is called *flow map* [Kraak & Ormeling 2003]. It should be considered that by convention flow symbols (e.g. arrows) represent only counts of items or amounts of goods moving between some places but not the routes of the movement. In Figure 2.7D there are many intersections among the flow symbols,

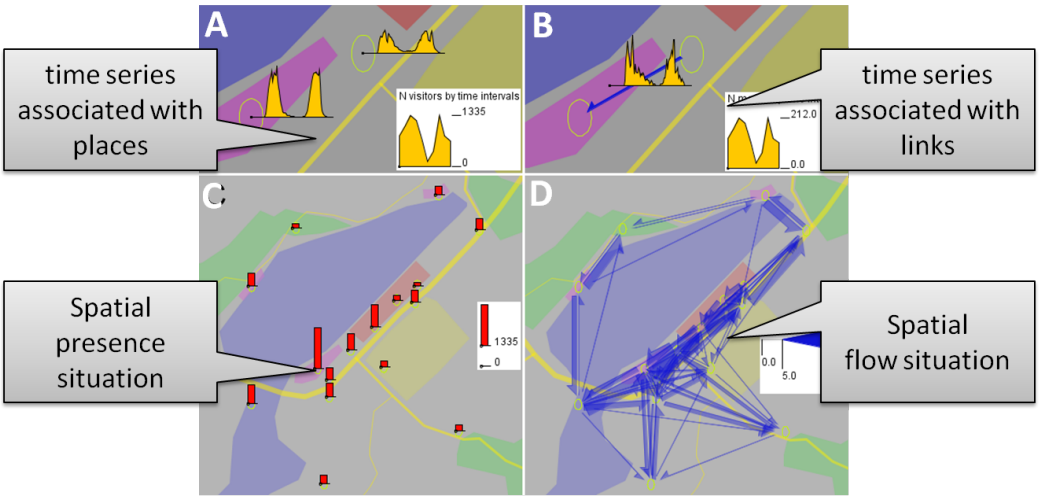

Figure 2.7: Views on Aggregated Episodic Movement Data: time series and spatial situations of presence and flows [Andrienko *et al.* 2012].

which clutter the display. This is a consequence of the discontinuity of the original trajectories, where consecutive recorded positions may be quite distant in space. For the brevity sake, we shall call spatial situations in terms of presence as *presence situations* and spatial situations in terms of flows as *flow situations* (and *undirected flow situations* derived from *undirected flow counts*).

2.3 Pedestrian Quantity Monitoring

Pedestrian monitoring provides the empirical data for pedestrian mobility analyses. Three major classes of traffic aspects are distinguished in our work and each provides different sensor technologies.

- **Microscopic** monitoring technologies return detailed individual movement data. These microscopic traffic aspects include trajectories and its attributes speed or movement direction. Possible sensor technologies include video surveillance systems or global satellite positioning systems (GPS).
- **Macroscopic** monitoring technologies do not measure values on individual movement but on groups. Examples for macroscopic values are quantity, density or average number of direction changes. Possible sensor technologies are manual quantity counting or Bluetooth tracking.
- **Infrastructural** traffic aspects give insights on usage of the provided infrastructure. Possible sensors are light switches, elevators and similar electronically equipped infrastructural elements.

The pedestrian quantities, which are the values focussed in this thesis, belong to the class of macroscopic values and thus macroscopic monitoring technologies are sufficient for our task. Depending on the application context, the pedestrian quantity is

either the number *counts of visits*, or the number of persons moving from one location to another during the considered time interval *counts of flows* (see previous section). Mostly, we refer to *pedestrian quantity* as the number of flows among two places. However, in the chapter on application scenarios (Chapter 6) we illustrate that with strong constraints our proposed methods may also be applied to *counts of visits*. In the following sections state-of-the-art technologies for quantity recording are presented. The presented methods include *video surveillance*, *manual counting* and radio frequency based methods such as *Global Positioning System* and *Bluetooth tracking*.

2.3.1 Video Surveillance

Naturally, observation of pedestrian mobility is a task for computer vision and using video surveillance seems to be a perfect method for empirical data collection at the first sight [Masoud *et al.* 2001]. Thus, methods are developed to extract trajectories [Fuentes & Velastin 2001] and the flow counts [Bertozzi *et al.* 2004] from video observations. Based on visual features (e.g. contour, texture or colour) [Seitner & Hanbury 2006] pedestrians are identified and recognized in a camera image (frame). In the continuous video stream the pedestrians need to be re-identified on every single frame. Major problems for video surveillance technology are:

- dependency on surrounding light,
- low image fidelity for far objects,
- occlusions by other objects, for example other pedestrians, trucks, or static objects.

Nevertheless, video surveillance systems perform well under controlled influences [Bertozzi *et al.* 2004]. Therefore, the camera systems often need to be mounted above the pedestrian stream. However, for industrial application in public space the need for special scaffoldings to carry the cameras are a main drawback, since often existing infrastructure may not be changed.

2.3.2 3D Laser Scans

Besides the classical video surveillance systems, 3D Laser scanner provide a data source for pedestrian tracking and pedestrian counting. Instead of colours, the pixels in a 3D Laser scan hold information of the distance between the camera and its surroundings. For example, the velodyne sensor sends omnidirectional laser rays and thus it stores the distances in a 3D image. An example of velodyne dataset, containing ground truth on pedestrian's movement is given in [Spinello *et al.* 2010, Spinello *et al.* 2011]. We depict a snapshot of this dataset in Figure 2.8. In literature numerous works tackled the task of pedestrian tracking, either in case of static [Teichman & Thrun 2012] or moving [Schulz *et al.* 2003] laser scanners. A combination with video surveillance is presented in [Kräußling *et al.* 2008]. However, the sensors are still very expensive and therefore not yet widely used.

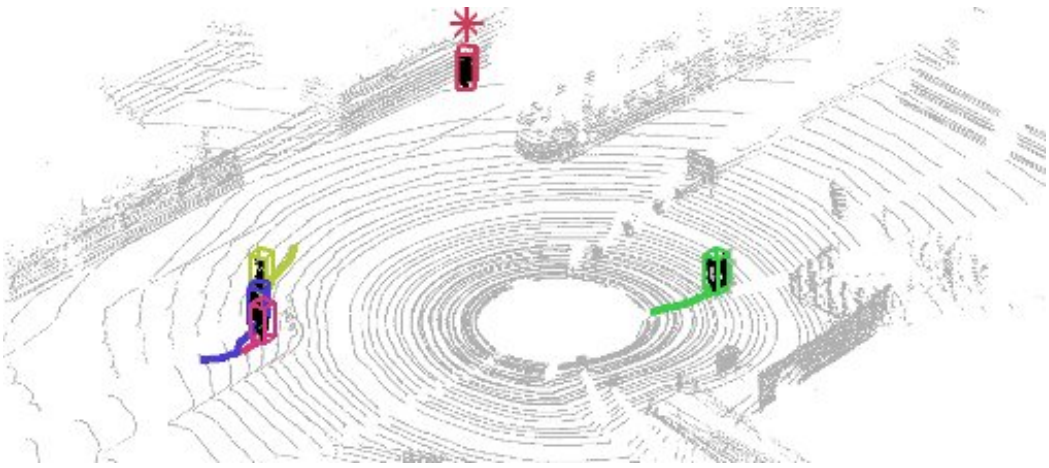

Figure 2.8: Annotated people and tracks in 3D Velodyne data [Spinello *et al.* 2011].

2.3.3 Manual Traffic Counting

Another method that overcomes the limitations of video surveillance is counting the number of pedestrian manually.

Figure 2.9: Smartphone application for manual pedestrian quantity monitoring.

In order to assist manual counting of flows and to simplify post-processing of measurements, we developed a smart phone application (Figure 2.9) which records clicks of the surveying person - each click represents the number of pedestrians passing by in a specified direction - along with its time-stamp. This enables an easy storage of the data in a database management system. Thus, the count of how many people passed at which time in which direction is collected (flow counts). To decrease the influence of the day of week on the measurements, the measurements can be repeated at three different days. As the number of "sensors" is limited, the locations need to be chosen carefully. In an early prototype, we encountered the problem of mixed directions; therefore we added visual hints to the smartphone application as well as to the map.

To distinguish directions, the colours red and green are used in our application. One major drawback of manual pedestrian counts is that they require lots of repetition and post-processing in order to be representative and accurate. For example, to be able to compare the empirical raw data of pedestrians at measurement locations, e.g. an average number of pedestrians for a complete week, post-processing is necessary: after merging the measurements, frequencies are weighted and aggregated according to the time interval and day when they were taken. As a result, every measured location has associated with it a number of passing pedestrians (flow counts) that may be compared against any other location. This is important for ranking locations within the closed environment. However, due to noisy input the manual pedestrian quantity counts tend to contain contradictions (i.e. there are more measurements of people who leave a junction than who enter it). The quantity estimation method we present in Chapter 3 resolves these automatically.

2.3.4 Radio Frequency Localization Techniques

In the last years, numerous radio frequency based tracking technologies have emerged. These technologies utilize one of the five possible methods [Fuller 2009]:

- **Proximity** is a cell based technology. Whenever a tracked device is close enough to a sensor to establish radio frequency communication (this is also called as entrance to the sensor's footprint) the location of the tracked device is mapped to this particular cell. Usage of multiple sensors offers the possibility to track movement.

- **Direction Finding (DF)** utilize rotatable sensor areas and utilize the proximity approach for a rotating sensor. Whenever a tracked device comes into the sensor area, the direction of its finding is known and position can be estimated by use of multiple sensors.

- **Angle Of Arrival (AOA)** computes the angle between the sensor and the device. The precision of (AOA) decreases with increasing distance due to the increasing influence of dispersion of the radio waves and their echo. AOA requires a free line of sight between the sensor and tracked device. Utilization of two sensors or more allows positioning of the tracked device by angulation methods (see [Fuller 2009]).

- **Timing and Phase of Arrival (TOA, POA)** computes the time difference the signal needs to traverse the distance between the sensor and tracked device. In case of multiple sensors, the device can compute distances to these sensors using a trilateration algorithm (see [Fuller 2009]). Larger distances improve precision of this localization technology. Time Of Arrival requires a precise synchronisation of the involved devices.

- **Radio Signal Strength (RSS)** utilizes the dependency among signal strength and distance. In theory, there exists an inverse proportional relation among these two values, i.e., the higher the distance the lower the signal strength. RSS

localization is easy to realise, no additional processing of received values is required (as it is for AOA and TOA). Similar to TOA, the position can be computed by trilateration algorithms.

- **Doppler** uses both signal strengths from the tracked device to the sensor and as well the way back from the sensor to the tracked device. The signal, which is sent from the tracked device, is compared to the one sent by the sensor. Thus, the impact of disturbances is reduced. The technology requires access to the data at the sensor and the tracked device.

Next, we describe two popular radio frequency tracking technologies. The first is called Global Positioning System (GPS) and utilizes POA (phase of arrival - described above) but depends on the line of sight. The second one, Bluetooth tracking, provides different methods (Proximity, DF, RSS) and receives increasing attention in mixed indoor/outdoor scenarios.

2.3.4.1 Global Positioning System

The Global Positioning System (GPS) was launched already in 1973 by the department of defense of the Unites States [National Academy of Engineering 2012]. Five years later the first NAVSTAR GPS satellite was sent to outer space. The system was ready for military use in 1988. In 1995, the GPS consisting of 24 satellites was also opened for civil usage. Nowadays, the system consists of 30 satellites cycling around the earth in six orbits. Per orbit there are four active satellites and one for redundancy, to compensate malfunctions. The altitude of the satellites is approximately 20,000 km. Every satellite rotates around the earth in 11 hours and 58 minutes. Given this satellite distribution, it is ensured that at every location on the surface of the earth at least four satellites are visible. International counterparts to Global Navigation Satellite Systems (GNSS) are for example [National Academy of Engineering 2012] GLONASS, Galileo, Compass/Beidou.

The system utilizes Phase of Arrival technologies (POA) [Fuller 2009]. The Global Positioning System depends on a free line of sight to the satellites. Thus, it is negatively influenced by reflections on building surfaces, refractions on trees, weather conditions (heavy snow) and by buildings. Due to the bad performance within buildings [Dedes & Dempster 2005] reliable indoor or combined indoor/outdoor scenarios are required to utilize other observation methods.

Thus, stationary placed Bluetooth scanners receive increasing attention in mobility monitoring. Next section introduces this technology.

2.3.4.2 Bluetooth Scanners

In GPS-less environments novel technologies are required for recording pedestrian mobility. Traditional approaches are surveys and video surveillance. Both are relatively expensive and difficult to realize. Surveys need to draw a representative sample of the pedestrians both in space and time. Video surveillance systems require additional scaffoldings that need to be integrated in existing architectures. Furthermore, quality of the video surveillances does depend on weather and light conditions.

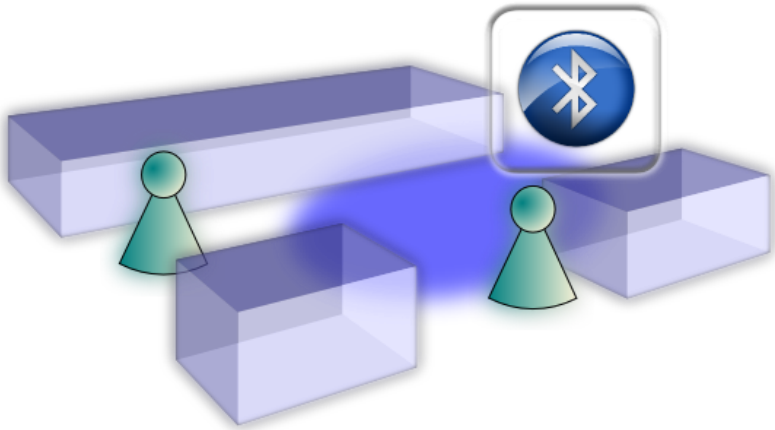

Figure 2.10: Bluetooth monitoring of pedestrian mobility. The blue circle represents the sensor's footprint.

The need for further robust passive localization technologies pushed the development of sensors that are capable to monitor people's movement. First choice is to track most popular digital gadgets: mobile phones and intercoms. Analysis of mobile network GSM (Global System for Mobile Communications) log files causes strong privacy objections [Giannotti & Pedreschi 2008]. Besides, Bluetooth technology is an emerging technology for monitoring tasks [Fuller 2009, Bruno & Delmastro 2003, Kolodziej & Hjelm 2006]. Recently evolved Bluetooth based mobility sensors have been used for event monitoring at a soccer match in France [Liebig & Kemloh Wagoum 2012] and a car race [Stange *et al.* 2011]. In these applications, a mesh of Bluetooth sensors has been placed at carefully selected indoor as well as outdoor locations. The work in [Liebig & Kemloh Wagoum 2012] extracts the route choices of the visitors and hands them to an agent based pedestrian microsimulation in order to extract microscopic movement values (compare Section 4.3.1).

The Bluetooth sensors (also scanners, transceivers or beacons) throughout this thesis are inspired by [Bruno & Delmastro 2003] and are assembled using a microcomputer and the three USB Bluetooth antennas. A Linux based software activates the antennas' inquiry mode and logs the (hashed [National Institute of Standards and Technology 2002]) MAC addresses of the detected devices[2]. Thus, the scan interval of approximately $10.24s$ [Woodings *et al.* 2001, Bruno & Delmastro 2003] is (theoretically) reduced to its third. This time span is required for frequency hopping and device discovery [Bluetooth SIG 2004]. However, in practice the antennas are not synchronous and data points are not recorded with a constant frequency. This problem was already addressed in the section on episodic movement data (Section 2.2.4). The recorded data log entries consists of:

```
[time stamp], [sensor ID], [sha256(MAC)], [signal strength] .
```

Two different antenna types which are used have a range of either $20m$ or $100m$. Nev-

[2]A similar software for Bluetooth Tracking is published by the University of Ghent at https://github.com/Rulus/Gyrid with GPL licence, last accessed 06/30/2012

ertheless, since the inquiry mode requires communication in both directions (from the sensor to the mobile device and vice versa) [Bluetooth SIG 2004] it not only depends on the sensor's antennas but also on the antenna in the mobile device (and its configuration). Thus, lower sensor ranges of about $20m$ occur. The recorded unique MAC addresses of the Bluetooth antennas of the mobile devices consist of six bytes. Three of them depend on the vendor [IEEE 2002] and provide valuable information for the analysis of collected data.

In [Andrienko *et al.* 2012] and previous Section 2.2.4 we addressed the formalization of recorded data. There, similarities to other episodic movement data are presented as well as methods for their aggregation and visual representation.

Besides event monitoring, also other successful indoor applications of Bluetooth scanners are described in literature. In [Pels *et al.* 2005] various scanners were placed at Dutch train stations to record transit travelers. Accurate locating and following of objects within complex facilities is as well an important research topic [Hallberg *et al.* 2003]. So far Bluetooth tracking is used to monitor a sample of visitors [Liebig & Kemloh Wagoum 2012, Andrienko *et al.* 2012, Hallberg *et al.* 2003] and extract their route choices [Liebig & Kemloh Wagoum 2012, Utsch & Liebig 2012]. The work presented in [Hagemann & Weinzerl 2008] uses Bluetooth tracking to track people among a public transportation network, whereas [Leitinger *et al.* 2010] gives a general overview on possibilities using Bluetooth tracking technology. In few works time-geography and movement patterns are addressed as well [Stange *et al.* 2011].

In this thesis we apply Bluetooth tracking for data acquisition in an indoor and a mixed indoor/outdoor closed environment and present novel methods for analysis of Bluetooth tracking data, see Chapters 4 and 6.

2.4 Pedestrian Mobility Models

Many approaches for modelling pedestrian mobility are described in literature. Similar to monitoring technologies, these approaches distinguish between microscopic and macroscopic aspects of mobility. Whereas microscopic models describe individual behaviour and provide trajectories for them, macroscopic models aim at modelling the moving population and use values as density, quantity or speed to characterize pedestrian flows.

2.4.1 Macroscopic Mobility Models

Macroscopic pedestrian models include physical models [Helbing 1997] which are based on the description of fluid and gas phenomena as well as specialized models on single macroscopic characteristics of movement (average speed, average density, quantity) per location. For the latter, a prominent example and focus of this thesis are the average daily traffic (ADT) maps, which denote traffic quantity (number of flows, compare Section 2.2.4) per location within a fixed time interval (e.g. per day). Next, we present both macroscopic views: (1) the all-embracing physical model, and (2) the ADT maps whose estimation is of central interest in this thesis, especially in Chapters 3 and 6.

2.4.1.1 Physical Model

Since the speed distribution of pedestrians is Gaussian distributed (compare Section 2.1) Henderson applied physical equations of gas motion to describe pedestrian movement [Henderson 1971, Henderson 1972]. Thus, the physical properties of the "crowd fluid" or "pedestrian gas" (e.g. viscosity) was estimated from empirical data. The density fluctuations in this pedestrian fluid cause shock waves [Helbing 1997]. However, in [Helbing 1997] doubts on the direct applicability of the physical gas equations are raised:

- If pedestrians interact, the impulse and the kinetic energy are usually not preserved. Thus, *Newton's Third law of motion* (actio=reactio) [Newton *et al.* 1803] is not applicable.
- Temperature of a crowd fluid cannot be matched directly, as it is the variance of the pedestrian speed.
- Pedestrian gases are not moving due to external *pressure*, but caused by the inner intention to move with a certain speed.
- Due to the various movement targets, separate flows in different directions occur and interact.
- Moreover pedestrian behaviour is *anisotropic*. For example distance or speed variances are higher in direction of motion than in perpendicular direction and pedestrians react more to events in front of them than to the ones behind them.

Therefore [Helbing 1997] proposes a new approach for fluid dynamical description of pedestrian flow. These equations are far beyond the context of this thesis, for a complete introduction to macroscopic mobility models we refer to the book [Helbing 1997].

2.4.1.2 Average Daily Traffic Map

Besides the physical model that aims to model all features of pedestrian mobility, models for particular features exist. For example *Annual Average Daily Traffic Maps* denote the number of pedestrian flows (compare previous Section 2.2.4) for street segments or corridors. Not all locations can be subject to empirical observations. Therefore the estimation of traffic quantities and the placement of the sensors are highly important research questions. In the following Chapter 3, our novel approaches incorporating movement patterns or movement pattern heuristics are presented after a brief survey on related approaches.

2.4.2 Microscopic Mobility Models

Microscopic pedestrian models describe individual behaviour and determine the macroscopic features of the pedestrian flow (e.g. the highly interesting *flow counts*, see Section 2.2.4) by aggregating simulated trajectories. Therefore, real-world pedestrians are represented by simulated *agents* and their behaviour is modeled based on the

empirical facts of pedestrian movement (see Section 2.1). Thus, the three-tier model for pedestrian motion [Hoogendoorn *et al.* 2002] (see Figure 2.1 in Section 2.1) became the fundamental concept for microscopic pedestrian modelling.

In an agent-based model, simulated pedestrians may be enriched with additional data, e.g. socio-demographic attributes, easily by linking representative data sources for example census.

However, there are mainly three different classes of models for pedestrian dynamic at the operational level: *Cellular Automata Models* [Blue & Adler 2001, Kirchner & Schadschneider 2002, Kretz & Schreckenberg 2006b], *Rule Based Models* [Thompson 1994, Galea *et al.* 2004, Korhonen *et al.* 2008, Raney & Nagel 2006] and *Force Based Models* [Molnár 1995, Yu *et al.* 2005, Chraibi *et al.* 2010]. Cellular automata have the advantage of being computationally efficient, but the resolution of the simulated geometry is limited by the size of the cells. Force based models usually operate on a continuous geometry, so they need more computations. For more about the advantages and disadvantages of the individual models we refer to [Schadschneider *et al.* 2009]. The microscopic models consist of many parameters and variables. Adjusting these variables to fit a model to empirical data is a hard task and requires multiple simulation runs [Kretz & Schreckenberg 2006a]. Therefore, main application field of microscopic models are evacuation simulations and short term traffic predictions, where the initial distribution of the pedestrian is well known.

Next, we describe two different microscopic pedestrian simulation approaches: (1) the *Generalized Centrifugal Force Model*, which operates in continuous space [Chraibi *et al.* 2011] and (2) the *Floor Field and Agent based Simulation* which utilizes cellular automaton [Kretz & Schreckenberg 2006b].

2.4.2.1 Generalized Centrifugal Force Model

If the operational level of the pedestrian walking is described by the Generalized Centrifugal Force Model (GCFM) [Chraibi *et al.* 2011] operating in continuous space, pedestrians are described with ellipses with velocity-dependent semi-axes. Faster ellipses (the spatial extent of agents that represent pedestrians) need more space in the moving direction. The motion is ruled by the social forces [Helbing & Molnár 1995, Molnár 1995]. At each simulation step the forces between the pedestrians and the obstacles (e.g. walls) are computed (compare Figure 2.11). Given an agent i with coordinates $\overrightarrow{R_i}$, the equation of motion is:

$$m_i \ddot{\overrightarrow{R_i}} = \overrightarrow{F_i} = \overrightarrow{F_i^{\text{drv}}} + \sum_{j \in \mathcal{N}_i} \overrightarrow{F_{ij}^{\text{rep}}} + \sum_{w \in \mathcal{W}_i} \overrightarrow{F_{iw}^{\text{rep}}}, \tag{2.2}$$

where $\overrightarrow{F_{ij}^{\text{rep}}}$ denotes the repulsive force from agent j acting on agent i, $\overrightarrow{F_{iw}^{\text{rep}}}$ is the repulsive force emerging from the obstacle w and $\overrightarrow{F_i^{\text{drv}}}$ is a driving force. m_i is the mass of agent i. $\ddot{\overrightarrow{R_i}}$ denotes the second time derivate of the location R_i (Newton's dot notation introduced in [Newton 1736]), therefore it represents the agents acceleration. $\mathcal{N}_i$ is the set of all agents that influences agent i and $\mathcal{W}_i$ the set of walls or borders that act on agent i. This model has been validated in corridors and bottlenecks using the fundamental diagram (which describes the dependency of average speed from

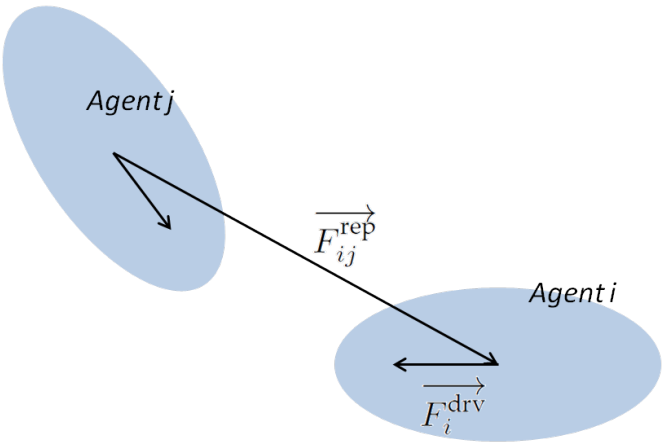

Figure 2.11: Generalized Centrifugal Force Model as introduced in [Chraibi *et al.* 2010]. Blue ellipses represent the spatial extent of moving agents.

the density [Helbing 1997]). The route choice for pedestrians can be done using a navigation field [Hartmann 2010, Guo & Huang 2011] or with a visibility graph. First approach is spread in the cellular automaton area. Continuous models usually work on a visibility graph, where the driving force of the simulated agent points towards a node of the graph. The strategies used here are usually the shortest path combined with the quickest path [Kretz 2009, Kirik *et al.* 2009, Heliövaara *et al.* 2011, Kemloh Wagoum & Seyfried 2011]. These strategies are in most of the cases validated using a visual assessment on some screenshots taken from the simulation. Some experiments have been conducted to determine pedestrians route choice using video surveillance, but only on simple scenarios, reducing the problem to an exit selection problem [Guo & Huang 2011, Heliövaara *et al.* 2011, Lo *et al.* 2006]. This is partially due to the fact that in complex facilities pedestrians have to be tracked across many rooms.

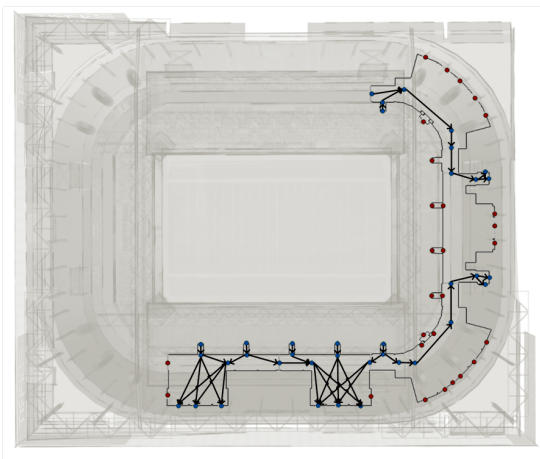

Figure 2.12: Example of a navigation graph generated from a section of a stadium considering which exits are closed [Liebig & Kemloh Wagoum 2012].

In the GCFM framework described here, pedestrians move from one decision area to the next one. A decision area is a place where the pedestrian decides which way to go or change the current destination. Ideally the decision areas are around the exits, which might be relevant for an evacuation scenario. The navigation network is automatically generated from the facility based on the inter-visibility of the exits, intermediate areas are inserted if needed. Visibility graphs can be constructed using different algorithms [De Berg *et al.* 2008, Höcker *et al.* 2010]. In the case of an evacuation scenario, the navigation graph can be limited to a visibility graph. This model has already been used to perform simulations of a multipurpose arena [Holl & Seyfried 2009, Seyfried *et al.* 2010]. A sample navigation graph for a section of a stadium is presented in Figure 2.12. Pedestrians are routed to the outside in this graph using four algorithms: the local shortest path, the global shortest path and a combination with the quickest path [Kemloh Wagoum & Seyfried 2011]. This example is suitable for an evacuation scenario where the pedestrians might prefer the shortest or quickest path to reach the outside. However, the approach is insufficient for normal day life situations, where the individual trips of pedestrians are subjected to other motivations. Some pedestrians might choose to go out the shortest way, whereas others might feel more comfortable walking along the promenade to get to some other points of interest. We therefore present a way to combine episodic movement data (namely Bluetooth Tracking) with the GCFM in the joint work [Liebig & Kemloh Wagoum 2012], also briefly reflected in Chapter 4. In Chapter 3 the open implementation of the GCFM[3] is considered as a baseline model for agent based quantity estimation.

2.4.2.2 Floor Field and Agent-based Simulation Tool

The *Floor Field and Agent-based Simulation Tool* is a cell based model of pedestrian movement. The model, introduced in [Kretz & Schreckenberg 2006b, Kretz & Schreckenberg 2006a] is discrete in space and time. The typical cell size is $40 \cdot 40cm^2$. The time slices are called *rounds* and are interpreted as one second. At maximum a cell can be occupied by one person at a time. The model makes use of three floor fields. The three floor fields contained by the model are the *static floor field*, the *dynamic floor field* and the *obstacle floor field*. They fulfil the following two tasks:

- The constant floor fields save calculation time, e.g. for distance calculations between exits and arbitrary positions.
- Dynamic floor fields that change over time can hold intermediate values and therefore transform long-ranged interactions into short-ranged ones.

The static floor field contains for each cell the distance of this particular cell to the exit. In case of multiple exits, there is a static floor field for each exit. The dynamic floor field is a vector field. If an agent moves from position $X_1 = (x_1, y_1)$ to $X_2 = (x_2, y_2)$ the dynamic field is updated by $(x_2 - x_1, y_2 - y_1)$ at position X_2. All values of both components in the dynamic floor field decay and diffuse with constant probabilities. In addition to the floor fields the agent movement is influenced by its properties.

[3] http://openpedsim.sourceforge.net/, last accessed 06/20/2012

- The agents are equipped with an inertia that prevents them from taking sharp turns. In contrast to the inertia introduced by Newtonian physics [Newton *et al.* 1803] it is not isotropic, but rewards acceleration and deceleration in the direction of motion.
- Furthermore, an agent can have a repulsive effect on other agents. If multiple agents compete for the same cell in a round, the model applies a friction parameter.
- The agents re-plan their decision for an exit in each round. To prevent the agents from jittering, the exit choice depends on the one made in the last round.

The simulation framework processes in each round the following steps [Kretz & Schreckenberg 2006a] (compare Algorithm 1).

Algorithm 1 FLOOR FIELD AND AGENT-BASED SIMULATION [KRETZ & SCHRECKENBERG 2006A]

```
for all agents do
    choose an exit
    choose a destination cell
end for
create sequence order O for the agents
for all agents ordered by O do
    execute movement step
end for
```

2.5 Mobility Patterns

As described in previous sections, pedestrians are no gases and the movement of individuals is no random walk but it is motivated by particular targets, route choices and influenced by other pedestrians (Chapter 1). Furthermore, pedestrians are not homogeneous in walking speed, space requirements and anatomy (Section 2.1). Thus, movement patterns occur in the crowd behaviour (Sections 2.1.3 and 2.1.4). In the following sections two representations of movement patterns are introduced: *Sequence Patterns* [Agrawal & Srikant 1995] and *Trajectory Patterns* [Giannotti *et al.* 2007]. Both patterns focus on different aspects of movement. Whereas *Sequence Patterns* are a representation for the location sequence traversed by trajectories (e.g. the *move sequences* which were defined in Section 2.2.4), *Trajectory Patterns* additionally model the travel times among the visited locations. For pedestrian quantity computations the temporal attributes can be neglected, thus in the next chapters we will use the term "pattern" in the sense of *sequence patterns*.

2.5.1 Sequence Patterns

The term *sequence pattern* has been introduced by [Agrawal & Srikant 1995] on arbitrary events. In order to transfer this notion to spatio-temporal context it is necessary to derive spatio-temporal events from movement, respectively trajectories. Having a

set of fixed discrete spatial locations, events may be either defined by the presence of a person at a specific location (compare *visit*, Definition 3) or its transition among two locations (*flow*, Definition 4) . The notion to model a trajectory as a sequence of separate activities was introduced by [Ellegård *et al.* 1977] and is based on the previously introduced space time geography [Hägerstrand 1970]. Thus, spatio-temporal events may be triggered by several spatio-temporal activities, for example *presence* at some location, *movement* at some location or *flow* among consecutive locations. Considering sequences of these spatio-temporal events results in *movement patterns* [Agrawal & Srikant 1995]. An example pattern representation may look as follows:

$$\left\{ \begin{array}{ll} T_1 & = l_1 \rightarrow l_3 \rightarrow l_5 \rightarrow l_6, \\ T_2 & = l_2 \rightarrow l_3 \rightarrow l_4, \\ T_3 & = l_4 \rightarrow l_5 \rightarrow l_1, \\ & \ldots \end{array} \right\} . \tag{2.3}$$

These kind of patterns will be applied in the next chapters as input data for pedestrian quantity estimation. In Chapter 4 we introduce methods to acquire and analyse movement patterns from observation data.

2.5.2 Trajectory Patterns

In addition to the sequence information contained by previously introduced *movement patterns* the *trajectory patterns* contain the transition times among the spatio-temporal events. Thus, a trajectory pattern has the structure:

Definition 5 (Trajectory Pattern) *A trajectory pattern is a sequence of successive spatio-temporal events e_i, e_j and their time intervals $\Delta t = t_j - t_i$. Therefore it is represented by sequences ($e_i \xrightarrow{\Delta t} e_j$) or, more general $e_0 \xrightarrow{\Delta t_0} e_1 \xrightarrow{\Delta t_1} \ldots \xrightarrow{\Delta t_{n-1}} e_n$.*

In [Giannotti *et al.* 2007] an algorithm is introduced for trajectory pattern retrieval from a set of GPS trajectories. This algorithm is two-fold: firstly, most popular spatial-regions are identified and secondly, the transitions among these regions are considered for inclusion in the trajectory pattern representation. However, neither GPS trajectories nor trajectory patterns are subject of this thesis, for further details on trajectory pattern mining in GPS logs we refer to [Giannotti *et al.* 2007].

2.6 Summary

For tackling the central challenge of this thesis, namely pedestrian quantity estimation in closed environments incorporating pattern knowledge, we presented the required fundamentals within this chapter. Thus, this chapter reflects the specifics of pedestrian movement and explains the natural occurrence of movement patterns. Furthermore the chapter describes state-of-the-art pedestrian tracking technologies (manual counting and video surveillance as well as radio frequency based: GPS and Bluetooth). In order to support the analysis of this observation data throughout the following chapters, methods for digital representation of pedestrian movement were introduced. Thus, we adopted the definition of *visits* at a location [Körner *et al.* 2010] and

additionally defined *flows* among multiple locations. For sparse (pedestrian) observation data, which carries the three uncertainties of *Continuity*, *Accuracy* and *Coverage*, we introduced the notion of *Episodic Movement Data* and presented methods for its spatio-temporal aggregation and visualisation on the map (firstly introduced in the joint publication [Andrienko *et al.* 2012]).

This thesis focuses on the estimation of the pedestrian quantities (macroscopic extent), therefore this chapter discussed pedestrian mobility models at varying granularities. These models comprise the macroscopic ones (which are focus of the thesis) and microscopic ones (i.e. agent based) which are utilized in the next chapter as a baseline model to the illustrative benchmark example. We introduced here models on spatio-temporal mobility patterns as well, since our approach makes use of pattern knowledge. The presented methods will again receive attention in the following chapters which focus on (1) our models for pedestrian quantity estimation incorporating movement patterns as well as (2) the analysis of these patterns in episodic pedestrian mobility observation data. (3) Finally, in the last chapter, we explain, driven by real-world scenarios, the application of presented concepts.

Chapter 3

Pedestrian Quantity Estimation Using Movement Patterns

"A big computer, a complex algorithm and a long time does not equal science."

—Robert C. Gentleman[1]

Contents

Previous chapters discussed the specifics of pedestrian movement and motivated the question for traffic quantities at locations. Empirical facts on movement behaviour have been highlighted as well as methods for digital data storage. We described how motivated individual mobility leads to patterns in mobility recordings and presented models for their representation. Furthermore, various pedestrian models have been discussed which describe different aspects of mobility (microscopic or macroscopic ones) based on preliminary assumptions. Utilizing this preliminary work, this chapter focuses on pedestrian quantity estimation using movement patterns. We present two contributed complementary regression approaches, Least Squares Regression (LSR) and Gaussian Process Regression (GPR), to tackle the task with different input data and constraints on movement. Whereas this chapter focuses on algorithmic

[1] Canadian statistician and bioinformatician, born 1959, originator of the R programming language [R Development Core Team 2009], at the Annual Meeting of the Statistical Society of Canada, Halifax, 2003

aspects of the quantity estimation and validates the presented approaches, following chapters address the pattern analysis in episodic mobility observation data and describe a software system for pedestrian mobility analysis as well as successful industrial applications of the presented methods.

3.1 Motivation

This chapter addresses the algorithmic part of the posed research questions (compare Chapter 1):

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

Details on the integration in real-world use cases are given in Chapters 5; Chapter 6 presents successful application scenarios.

Estimation of pedestrian quantities is an important question to various applications. Though naturally occurring movement patterns are not incorporated into existing approaches, these patterns contain valuable information on spatial correlations among traffic quantities.

Consider for example an average daily traffic (ADT) prediction problem with traffic networks consisting of only a single junction. As shown in Figure 3.1, 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 along the main corridor. Assume further to have a frequency sensor (compare previous chapter for a survey of suitable technologies) placed in the main corridor which measures a known amount of people within considered time interval. Under these circumstances, existing traffic volume estimation methods do not take into account the expert knowledge and thus may not effectively provide accurate estimations for the side corridor [Liebig *et al.* 2012b] (see Section 3.5). Throughout this chapter, we assume the movement patterns are known (explicitly or by heuristics). In the next chapter, we address the task to record and analyse these patterns from observation data. The two novel quantity estimation methods presented in this chapter utilize linear and Gaussian process regression. Furthermore, the models utilize:

- heuristics on route choices or movement pattern,
- the floor plan topology and
- readings of few pedestrian quantity sensors.

Since in real-world scenarios floor plan sketches are often available and few sensor readings can be obtained at low expenses, our methods meet the requirements for applications in real-world (compare Section 1.1).

After introduction of the proposed methods, this chapter proceeds with the performance analysis of the methods. At the end of this chapter, both algorithms (LSR

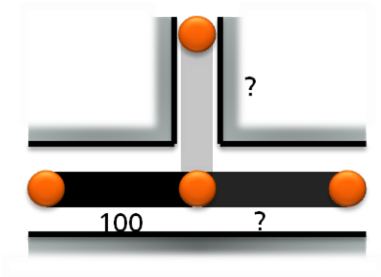

Figure 3.1: T-Junction example. Pedestrian quantity is measured in the main corridor (to the left). At the junction a small corridor intersects and an expert knows that it is most likely for pedestrians to continue their walk straight on.

for use with heuristics on mobility patterns and GPR for applications where mobility patterns are explicitly known) are validated against state-of-the-art approaches for the T-Junction benchmark (see Section 3.8).

3.2 Preliminary Definitions

Before methods for pedestrian quantity estimation can be addressed, preliminary definitions of the domain are required. These foundations are given in this section.

This theses focuses on quantity estimation in *closed environments*, we define them as follows:

Definition 6 (Closed Environments) *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 which connect them with their surroundings. Prominent examples are train stations, terminals, shops, shopping malls, parks, as well as zoological gardens. Only in closed environments pedestrian quantity estimation can be analysed without any unexpected influences (for example the locations with arbitrary stay times from urban environments e.g. living houses and points of interest), thus Kirchhoff's law [Kirchhoff 1845] is fulfilled for the quantity of people within a fixed time period (number of incoming people equals the number of outgoing ones).

Based on the previous definition the representation of the walkable area by networks is reasonable, since pedestrians may only enter and exit the area via dedicated exits (entry and exit points). Edges of the so-called *traffic network* represent the corridors and paths, whereas junctions are represented by vertices.

Definition 7 (Traffic Network) *A traffic network* $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ *is an abstraction of a floor plan by an undirected graph which consists of*

- *a finite set of vertices* $\mathbf{V}$ *which represent junctions, and*
- *connecting edges* $\mathbf{E} \subseteq \mathbf{V} \times \mathbf{V}$ *which represent corridors.*

Algorithms for traffic network construction from a given floor plan sketch are so-called *triangulation methods*. Triangulation methods are closely related to tessellations as they may become mapped to each other: every vertex of a triangulation becomes the handle of a surrounding polygon. These polygons of the tessellation are adjacent if and only if the vertices of the triangulation were connected. Tessellations partition the walkable manifold in discrete compounds (compare Section 2.2.4.1), whereas triangulations are a graph representation of the tessellation topology.

A commonly applied automatic method for traffic network construction is the Delaunay triangulation [Delone 1934], its tessellation counterpart is the Voronoi tessellation [Dirichlet 1850, Voronoï 1908] (introduced in Section 2.2.4.1). The Delaunay triangulation algorithm constructs a mesh of adjacent triangles among a given set of vertices. The process which automatically derives a traffic network from a floor plan is described in [Demyen & Buro 2006]. Briefly summarized, (1) vertices are drawn uniformly at random from the walkable area of the floor plan sketch; (2) the Delaunay triangulation among these vertices returns the traffic network.

Another possible method for traffic network construction from a given floor plan is the manual creation of vertices at junctions and connecting edges among the corridors.

This method does not depend on exact geometric representation but may also handle floor plan sketches, if they provide the topology of the walkable area. An example of the traffic network for Hofheim central station is depicted in Figure 3.2.

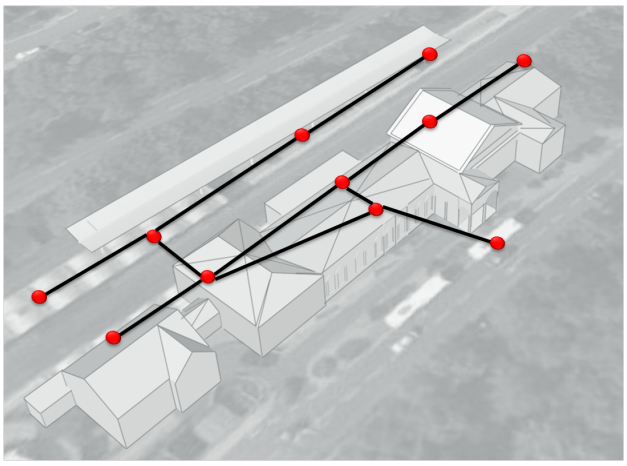

Figure 3.2: Example of a Traffic Network for a Closed Environment (Hofheim central Station). Black edges represent corridors, red vertices mark junctions. [Liebig *et al.* 2012b].

Given the *traffic network* of a *closed environment* the *pedestrian quantity* $q(e, \Delta t)$ for an edge $e \in \mathbf{E}$, $e = (v_i, v_j)$ is the undirected *flow count*,

$$q(e, \Delta t) \; := \; NM(e, \Delta t) \; = \; NM_u(v_i, v_j, \Delta t) \; ,$$

(compare Section 2.2.4.1) among the locations v_i and v_j within Δt. Without loss of generality, we assume to consider the same Δt at all edges and thus consider the *flow situation*[2] (compare Section 2.2.4.1). Therefore, we assign:

$$q(e) := q(e, \Delta t) \; .$$

Hence, the time interval for the *flow situation*, Δt, is considered to be the same over all edges in the *traffic network* so the time dynamic during Δt is neglected.
Given this explanation, we derive the *pedestrian quantity* $q(e)$ as follows.

Definition 8 (Pedestrian Quantity) *$q(e)$ at an edge $e \in \mathbf{E}, e = (v_i, v_j)$ in a traffic network $\mathcal{G} = (\mathbf{V}, \mathbf{E})$, within time interval Δt, are the number of persons which were passing the path segment, represented by e, within Δt from one vertex v_i to v_j or vice versa.*

Regarding the duality among a triangulation and a tessellation (described above) the pedestrian quantity at edge e can similarly been considered as the number of persons crossing the boundary between two adjacent spatial polygons. The same notion of pedestrian quantity applies directly for video surveillance or manual counting recording technology (described in previous Chapter 2).

[2]Despite mostly applying this definition, note that in Section 6.4.4 we are exploring the direct usage of the *number of visits*: $q(e, \Delta t) \; := \; NV(e, \Delta t)$.

Definition 9 (Pedestrian Quantity Measurements) *The pedestrian quantity measurements* $\mathbf{y} = \{y_s\}$ *denote the pedestrian quantity at a bounded number of sensor locations* $e_s \in E_S, E_S \subseteq \mathbf{E}$ *among a traffic network* $(\mathbf{V}, \mathbf{E})$*, with*

- $y_s := \{NM_u(e_s, \Delta t) : s = 1, \ldots, S\}$ *(where* $NM_u(e_i, \Delta t)$ *was defined in Section 2.2.4.1 as undirected number of moves along the edge* e_i*) and* $S = |E_S|$ *.*

These definitions of *pedestrian quantity* and *pedestrian quantity measurements* are based on the notion of spatial events (*moves* among two location, as introduced in previous Chapter in Definition 4) caused by individual *trajectories* (Definition 2) and its aggregate (Section 2.2.4.1). Thus, this abstraction of pedestrian movement holds the requirements for spatial models posed in [Hägerstrand 1974]:

- The model is related to the things happening in reality.
- It is easy to move between microscopic and macroscopic levels without loss of its relation.

After defining the notions regarding *traffic network*, *pedestrian quantity* and *pedestrian quantity measure*, we proceed with the definitions regarding *movement patterns*.

As described in previous discussions, pedestrian mobility causes naturally occurring movement patterns (compare Chapters 1 and 2). These patterns are input to our quantity estimation methods, besides *traffic network* and *pedestrian quantity measurements*. Thus, we distinguish between *movement patterns* (based on sequence patterns [Agrawal & Srikant 1995], see Section 2.5.1) and *movement pattern heuristics*.

Definition 10 (Movement Pattern) *A movement pattern* $\mathcal{T} \in \mathbf{T}$ *is represented by a sequence of co-visited locations* $\{\mathcal{T} = T_1 \rightarrow T_2 \rightarrow \ldots | T_i \in \boldsymbol{E}\}$ *among the traffic network* $\mathcal{G}(\boldsymbol{V}, \boldsymbol{E})$*.*

The likeliness of a particular movement pattern to occur is influenced by the pedestrian route choice and depends on the characteristics of the pattern, e.g. its attraction, safety or detour. Thus, a movement pattern heuristic defines a partial order among movement patterns based on its characteristic.

Definition 11 (Movement Pattern Heuristic) *A movement pattern heuristic is a preorder on movement patterns* $\mathcal{T}$ *based on a characteristic function* $f : \mathcal{T} \longrightarrow \mathbb{R}$*, which satisfies:*

- $\mathcal{T}_i \succ \mathcal{T}_j \iff \mathcal{T}_i$ *is more likely than* $\mathcal{T}_j$ *,*
- $\mathcal{T}_i \succ \mathcal{T}_j \iff f(\mathcal{T}_i) \leq f(\mathcal{T}_j)$ *,*
- $\mathcal{T}_i \succ \mathcal{T}_i$ *(reflexivity),*
- $\mathcal{T}_i \succ \mathcal{T}_j \wedge \mathcal{T}_j \succ \mathcal{T}_k \Longrightarrow \mathcal{T}_i \succ \mathcal{T}_k$ *(transitivity).*

The *Movement Pattern Heuristic* is thereforecompletely described by the characteristic function f. An example for a movement pattern heuristic is the preference for the shortest route in train stations, which was empirically justified in [Li *et al.* 2008].

This example also demonstrates why the heuristic is a preorder as the antisymmetry property ($\mathcal{T}_i \succ \mathcal{T}_j \wedge \mathcal{T}_j \succ \mathcal{T}_i \Longrightarrow \mathcal{T}_i = \mathcal{T}_j$) is not guaranteed due to the unconstrained mapping of $f(\mathbf{T})$.

This section specified the domain for the pedestrian quantity estimation. We present in the next section methods for the crucial problem of traffic volume estimation from sparse measurement data. This was already addressed in literature, although mostly for vehicular traffic [May *et al.* 2008]. Thus, we next provide an overview of existing methods from different research areas (operation research and data mining). Afterwards, in proceeding sections, our contributed methods are presented and compared with state-of-the-art related approaches.

3.3 Problem Statement

Utilizing the preliminary definitions regarding traffic networks, pedestrian quantities and their measurements, we examine the problem of appropriately determining the traffic quantities at unmeasured edges in a traffic network. This problem can be formulated in the following way: Given a graph $\mathcal{G}(\mathbf{V}, \mathbf{E})$ and for some edges $E_S \in \mathbf{E}$ their undirected pedestrian quantity measurements $\mathbf{y} = \{y_s := NM_u(e_s, \Delta t) : s = 1, \ldots, S\}$ find the undirected quantities $\mathbf{x}$ for the unobserved edges $E_U = \mathbf{E} \setminus E_S$. In literature, compare [Zhao & Park 2004, Gong & Wang 2002, Neumann *et al.* 2009], a common criteria to evaluate the performance of a traffic quantity estimator is the computation of the mean absolute error (MAE):

$$MAE := \frac{\sum_{j=1,\ldots,|E_U|} abs(x_j - NM_u(e_j, \Delta t))}{|E_U|} \text{, with} (x_j \in \mathbf{x}, e_j \in E_U) \ . \tag{3.1}$$

In the next section, we present approaches from the literature for solving the described problem.

3.4 Related Approaches

Estimation of traffic quantities at unmeasured locations implies the computation of missing values from the input data. This problem is known as the *imputation problem* in statistics. In general, every missing value is filled with a plausible value, e.g., one particular sensor reading or the global mean of the sensor readings. More complex imputation method are regression algorithms which utilize functional dependencies among the variables to fill missing values. For spatial data, these dependencies occur naturally in spatially distributed phenomena due to a statistical effect called spatial auto-correlation. This spatial auto-correlation is expressed by *Tobler's law* as follows: "Everything is related to everything else, but near things are more related than distant things" [Tobler 1970].

Regression methods became a major topic in data analysis and data mining [Witten & Frank 2005]. Various models for the functional dependencies among the values of a domain were created. This thesis focuses on regression approaches for the presented

quantity estimation problem (for an introduction to regression models we refer to [Witten & Frank 2005]).

In particular, we describe here the latest developments and commonly used algorithms for traffic quantity estimation though some of them reflect *Tobler's Law* in their presumptions on traffic flow, so far none of the presented related approaches incorporates movement patterns. These patterns contain valuable information and after the literature survey, we introduce our approaches that incorporate either heuristics or knowledge on movement patterns. In general, the characteristic properties of pedestrian flow make the application of existing regression approaches for fluids in pipe networks [Tanyimboh & Templeman 1993, Yassin-Kassab *et al.* 1999] unfeasible, compare Section 2.4.1.1 and [Helbing 1997] for differences of pedestrians and fluids.

Existing literature distinguishes between average daily traffic (ADT) estimation and average annual daily traffic (AADT) estimation [FDOT 2012]. 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. A naïve approach for AADT estimation is the utilization of ordinary linear regression (OLR) method [Zhao & Park 2004]. Street segment (edge) attributes (e.g. number of lanes or function classes) are multiplied by weights which become subject for the regression.

Definition 12 (Ordinary Linear Regression) *The Ordinary Linear Regression models a missing variable x_i as follows:*

$$x_i = \beta_0 + \sum_{k=1}^{p} \beta_k y_{ik} + \varepsilon_i \ . \tag{3.2}$$

Where β_k is the k-th parameter for the OLR model, y_{ik} is the independent variable of the k-th parameter at location i, ε_i is the error term at location i, and p is the number of parameters [Zhao & Park 2004].

Improvements of this technique were achieved by respecting the geographical space by usage of geographical weighted regression (GWR) [Zhao & Park 2004] and by the application of k-nearest neighbor approaches (kNN) [Gong & Wang 2002].

Definition 13 (Geographical Weighted Regression) *The Geographical Weighted Regression models a missing variable x_i similar to OLR by substituting local parameters for global ones as follows:*

$$x_i = \beta_{i0} + \sum_{k=1}^{p} \beta_{ik} y_{ik} + \varepsilon_i \ . \tag{3.3}$$

Where β_{ik} is the kth parameter at location i for the GWR model, y_{ik} is the independent variable of the kth parameter at location i, ε_i is the error term at location i and p is the number of parameters [Zhao & Park 2004].

Definition 14 (k-Nearest Neighbour Regression) *The k-Nearest Neighbour regression uses the value of the k closest data points $N_k(x_i)$ for estimation of the missing ones $x_i \in \mathbf{x}$.*

Closeness d is measured in feature space.

$$
\begin{aligned}
&N_k(x_i) \subseteq Y, |N_k(x_i)| = k, \\
&\nexists y_j : y_j \in Y \wedge y_j \notin N_k(x_i) \textit{ having } d(y_j, x_j) < d(y_n, x_i), y_n \in N_k(x_i), n = 1 \ldots k \\
&y_i = f(N_k(x_i))
\end{aligned} \tag{3.4}
$$

In [Lam *et al.* 2006] the GWR method is compared to the kNN approach and the Gaussian maximum likelihood (GML). The kNN achieved best results. Additionally, the AADT prediction of k-Nearest Neighbour for a particular location is improved by weighting measurements by their temporal distance to the prediction time. This approach showed better results than the application of Gaussian maximum likelihood (GML) approaches for weighting the historical data points.

Definition 15 (Gaussian Maximum Likelihood) *The Gaussian maximum likelihood approach to the AADT problem assumes Gaussian distribution of the flow counts $y(t, l)$. Among two discrete time intervals Δt. Also the change of the flow count $dy(t, l) = y(t, l) - y(t - \Delta t, l)$ is assumed to be Gaussian distributed.*

$$
\begin{aligned}
y(t, l) &\sim \mathcal{N}(\mu_{y,t}, \sigma^2_{y,t}) \\
dy(t, l) &\sim \mathcal{N}(\mu_{dy,t}, \sigma^2_{dy,t})
\end{aligned} \tag{3.5}
$$

According to [Lam et al. 2006] the closed form estimate for $y(t, l)$ is then:

$$
y(t, l) = \frac{\sigma^2_{y,t}(\mu_{dy,t} + y(t - \Delta t, l)) + \sigma^2_{dy,t}\mu_{dy,t}}{\sigma^2_{y,t} + \sigma^2_{dy,t}} \ . \tag{3.6}
$$

However, this GML method focuses on the dynamics of traffic flow, whereas our thesis tackles the estimation of quantities for unmeasured locations. Addressing a similar problem, recent improvements to the k-Nearest Neighbour non parametric regression were made in [May *et al.* 2008]. In [May *et al.* 2008] the geographic coordinates are a subcomponent of the attribute space. Furthermore, this work addresses the ADT estimation problem 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 [May *et al.* 2008]. Their proposed algorithm is a spatial k-nearest neighbour (S-kNN) approach that incorporates geometric distances in the kNN attribute vector for estimation of an unknown edge. The closer a measured edge is to an unmeasured one (in attribute space), the higher its impact. This is similar to the Kriging approach described in [Wang & Kockelmann 2009] but goes beyond it, since for the latter prediction only the k-nearest neighbours were used.

3.5 Comparison of State-of-the-art Quantity Estimation Approaches

Traffic volume estimation is a natural task in street-based traffic analysis systems and has important applications, e.g., quality-of-service evaluation, location evaluation or risk analysis. Estimation of pedestrian volumes received recent attention as it offers vast possibilities to extract motivations of people's movement and helps to improve and to plan provided infrastructures. Compare Section 3.3 for the formal problem statement. Distinguishing from the general large scale traffic volume estimation, pedestrian mobility in closed environments (e.g., train stations) holds the additional property that individuals are likely to have homogeneous movement preferences (Section 2.1.3). Thus, movement patterns exist and have high impact on the distributions of traffic among the small networks. The requirements posed to our quantity estimation algorithm are therefore (compare Section 1.1):

- exclusively based on floor plan sketches and public available data sources,
- adjustable by quantity measurements,
- incorporating knowledge on movement pattern.

This section evaluates the performance of existing traffic network based approaches (pedestrian simulation, Spatial k-Nearest Neighbour and Gaussian Process Regression)[3] for applicability in a highlighting closed environment scenario. None of the state-of-the-art approaches incorporates knowledge on movement preferences. Therefore all of the hereby tested methods do not hold the third requirement. The consequences are shown in a small striking example. Other, more complex scenarios, are described and compared (in terms of MAE, see Section 3.3) in the validation sections of this thesis, respectively, in the chapter on application scenarios, Chapter 6.

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 on movement patterns is thus essential to address the above challenges. However this has not been well investigated in existing work. Consider for example an average daily traffic (ADT) prediction problem with a traffic network consisting of only one junction. As shown in Figure 3.3, a T-Junction occurs in a wide corridor that goes straight ahead. At the junction a small corridor intersects and an expert knows that it is most likely for persons to continue their walk straight ahead along 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 [Liebig *et al.* 2012b].

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. In detail the experiments were carried out as follows.

[3]We omit Geographically Weighted Regression (GWR), Gaussian Mixture Models (GML) and Spatial k-Nearest Neighbour (S-kNN) which perform worse than hereby considered methods, compare Section 3.4.

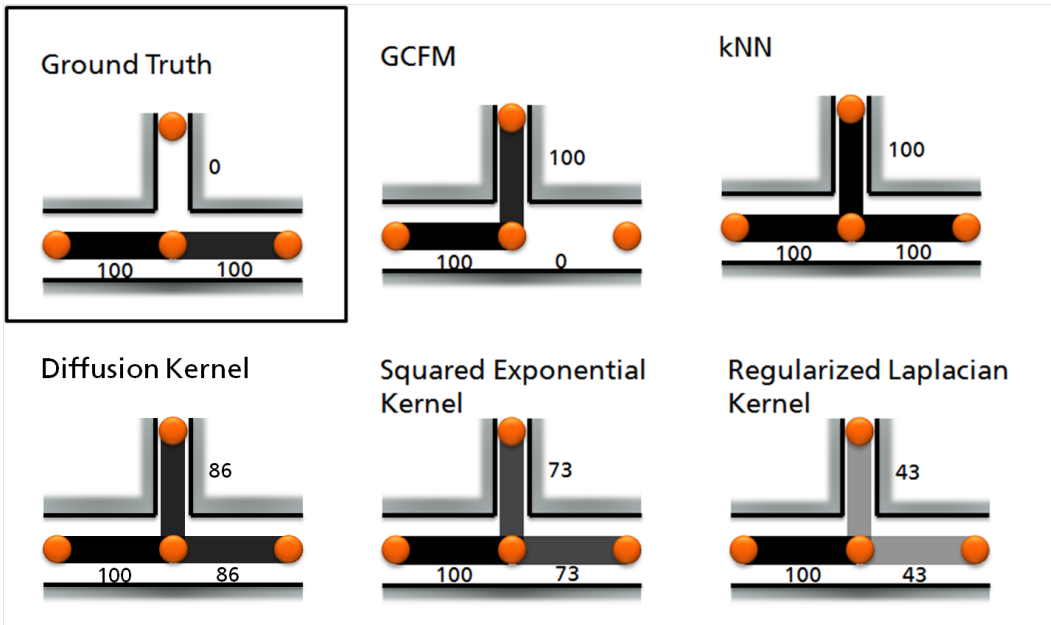

Figure 3.3: T-Junction example. Left corridor is measured, quantity of other corridors is unknown. Numbers denote relative frequencies per edge. Related Approaches not incorporating Movement Patterns are applied: pedestrian simulation (GCFM), Spatial k-Nearest Neighbour Method S-kNN and Gaussian Process Regression using different kernel functions (lower row).

- The **Pedestrian Simulation** has been performed with the state-of-the-art social force model for pedestrian movement. This so called *Generalized Centrifugal Force Model (GCFM)* [Chraibi *et al.* 2011] is an agent based microscopic simulation, and has already been introduced in Section 2.4.2.1. An open source implementation of the GCFM for pedestrian simulation can be freely obtained on the Internet[4]. In contrast to the other discussed approaches in this chapter, the GCFM does not only depend on a traffic network which represents the topological floor plan information, but requires a more detailed geometric representation of the walkable area. In this floor field every simulated pedestrian (*agent*) has ellipsoidical space requirements. Depending on the agent's speed the ellipse becomes larger in the direction of motion. Every agent is influenced by a couple of forces. These divide into two classes:

  - *repulsive forces* prevent agents from hitting obstacles (walls or neighbouring agents),
  - *driving forces* cause the motion of the agent in its desired target direction.

  So far the open source implementation does not allow for manually defined target choices of the agents. As this software was implemented for evacuation simulation (where the agents need to decide individually for their exits) the lack of this feature was on purpose. However, in the considered T-Junction example

[4]http://openpedsim.sourceforge.net/, last accessed 06/20/2012

we bypass this by initiating multiple agents in the left side of the main corridor and just define two exits: one in the main corridor and another one in the side corridor.

- The **Spatial k-Nearest Neighbour algorithm (S-kNN)** is a graph based macroscopic approach. It applies a distance function in feature space among the street segments to decide for closeness. The k closest (in feature space) measured edges are considered for computation of the weighted average for an unknown one. In the T-Junction example just one edge is given and all edges have contact at the junction. Thus, any arbitrary chosen $k \geq 1$ will generate the same k-neighbourhoods which contain the only given street segment. As all edges contact each other, their distance is zero. This distance may be also computed very fast, using the bounding-box based partial evaluation scheme from [May *et al.* 2008]. Thus, all edges receive the same quantity (Figure 3.3 in the upper right picture) which has previously been monitored at the given edge. However, to prevent this behaviour, there are possibilities to adjust the distance function in order to obtain different results. One possibility would be to calculate distances among the edges not just in geographical space but in a vector feature space. Thus, besides its spatial extent every edge is equipped with a list of features. The distance function, used to decide for the distance weighting in the k-Nearest Neighbour approach can be defined on these features. Thus, more similar edges in feature space result in more equal quantity values. In detail, application of such an extension to the corridor type could be: Addition of a corridor class attribute to every edge, which numerically encodes the type *main corridor* or *side corridor*. With well chosen distance function, this may prevent the S-kNN from estimation of any frequency for the small corridor. However, even with a large list of features and an optimisation on their weights within the quantity estimation function of the S-kNN no movement pattern knowledge is represented in the S-kNN. Since the algorithm does not compute new values but performs weighted averaging, the estimated quantities may not reach unobserved quantities but remain within the range of the observed ones.

- We also compared **Gaussian Process Regression** for pedestrian quantity estimation. Thus, we utilized the following graph kernels: *Regularized Laplacian* [Smola & Kondor 2003], *Squared Exponential* (as recently proposed in [Selby & Kockelman 2013]) and the *Diffusion Kernel* [Kondor & Lafferty 2002]. In detail the Gaussian Process Regression model is explained in Section 3.7.

The experimental results in the T-Junction example (Figure 3.3) reveal that existing methods which do not reflect expert knowledge perform badly. Therefore novel approaches for pedestrian quantity estimation, taking into account pedestrian movement preferences, are required. Next sections describe our contributed pedestrian quantity estimation methods. At the end of the chapter we revive the T-Junction example and apply our novel approaches with great success.

3.6 Approach with Pattern Heuristic: LSR

Usually assumptions on the preferred route choices of the people are given by experience of some domain expert. In train stations for example, the assumption exists that most people prefer the quickest path for their movement [Rindsfüser 2005], and the group of persons who behave differently is insignificantly small. This commonly made assumption was justified by empirical laser beam measurements in [Li *et al.* 2008] and is well established in pedestrian traffic analysis.

In this section, we present one of our methods[5] for pedestrian quantity estimation which incorporates a site's traffic network, quantity measurements, as well as movement pattern heuristics.

Empirical recordings of pedestrian flow counts denote quantities $\mathbf{y}$ at pre-selected edges (compare Definition 9, Section 3.2). According to this definition, the measurement y_s holds the undirected quantity $q(e_s)$ at a particular edge e_s in a time interval Δt which is defined by the sum of all pedestrians walking along the edge during Δt:

$$\mathbf{y} = \{y_s := NM_u(e_s, \Delta t) : s = 1, \ldots, S\} \ . \tag{3.7}$$

However, the number of possible undirected, acyclic paths in the traffic network which contain e_s is bounded. These paths through a closed environment can be expressed by sequences of traversed edges in *Movement Patterns* (Definition 10) under the following additional conditions.

Definition 16 (Path) *A path through a traffic network $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ is a movement pattern $\{\mathcal{T} = T_1 \rightarrow T_2 \rightarrow \ldots | T_i \in \boldsymbol{E}\}$ with:*

- $e_i \rightarrow e_j \in \mathcal{T} \Longrightarrow e_i = (v_1, v_2) \wedge e_j = (v_2, v_3)$ *(Adjacency),*
- $|\mathbf{v}| = |\mathcal{T}| + 1$, *with* $\mathbf{v} = \{v \in e_i \forall e_i \in \mathcal{T}\}$ *(Acyclic),*
- e_1 *and* $e_{|\mathcal{T}|}$ *connect the closed environment* $\mathcal{G}$ *with its surroundings.*

Moreover, the quantity $q(e_s)$ is also denoted by the number of pedestrians who use a path that contains e_s (compare Figure 3.4):

$$q(e_s) = \sum_{\mathcal{T}: e_s \in \mathcal{T}} q(\mathcal{T}) \ . \tag{3.8}$$

[5] **published with a major contribution of the author in:**
T. Liebig and Z. Xu. *Pedestrian monitoring system for indoor billboard evaluation*. Journal of Applied Operational Research, vol. 4, pages 28–36, 2012.
T. Liebig. *Trajectory Regression Model for Indoor Pedestrian Flow Analysis on Billboard Evaluation*. In Proc. of the Third International Conference on Applied Operation Research - ICAOR'11, pages 289–300. Tadbir Operational Research Group Ltd., 2011.

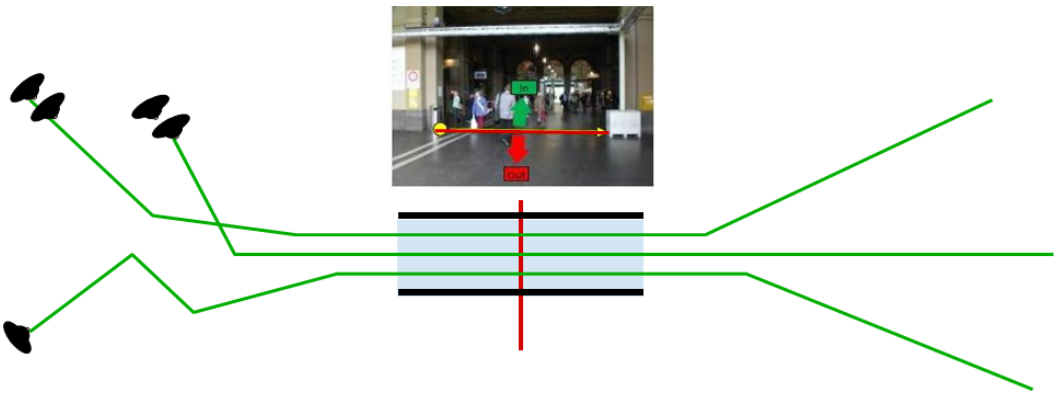

Figure 3.4: Three paths taken by five persons (marked with black symbols in the left) passing a corridor, causing a frequency of five persons within the observation time interval Δt.

Therefore our approach consists of three consecutive steps:

- (1) The enumeration of the set of all valid (acyclic) paths passing the site.
- (2) The computation of path frequencies such that the differences at the measured locations becomes minimal.
- (3) The estimation of the quantity of pedestrians for all edges in the traffic network.

The three generated output variables, edge quantities, path-frequencies and the enumerated set of paths provide the desired information to the billboard campaign evaluation scenario introduced in Chapter 1 and revised in Chapters 5 and 6.

Previous empirical studies reveal that people prefer routes with the minimal detour most, when walking from a given start to a target [Rindsfüser 2005]. Thus, the assumption that most trajectories are acyclic is reasonable for public buildings. For the special case of train stations [Li *et al.* 2008] justifies the assumption that most people choose the shortest route. Empirical observation we made during the application also supports this assumption. We consider cyclic paths to be irrelevant for our study. These assumptions describe some differences among pedestrian flow and flow of fluids or gases, for more differences we refer to Section 2.4.1.1. These characteristic properties of pedestrian flow make the application of existing regression approaches for fluids in pipe networks [Tanyimboh & Templeman 1993, Yassin-Kassab *et al.* 1999] unfeasible. By application of a characteristic function for each path, we construct for all paths its corresponding binary representation (vector) of constant length. Every position in the resulting vector relates to an edge.

$$A = (a_{ji})_{0 \le j \le |\mathcal{T}|}^{0 \le i \le |E|}$$
$$a_{ji} = \begin{cases} 1 & \text{if } e_i \in \mathcal{T}_j, \\ 0 & \text{otherwise} \end{cases} \tag{3.9}$$

The vector equals one at a position if the related edge is element of the path and zero otherwise. The resulting vectors for every previously enumerated (plausible) path is aggregated in a matrix A. By application of this characteristic matrix, the edge frequencies can be expressed by the matrix multiplication of this matrix with the path frequencies.

$$\begin{aligned} y_{e_i} &= \sum_{1 \leq j \leq |\mathcal{T}|} a_{ji} \cdot y_{\mathcal{T}_j} \\ y &= A^T \times y_{\mathcal{T}} \end{aligned} \tag{3.10}$$

The program which solves step (2) derives directly as:

$$\begin{aligned} & y^*_{\mathcal{T}} = \arg\min \left(\left\| A^T \times y_{\mathcal{T}} - y_{e_s} \right\| + \sum_{1 \leq j \leq |\mathcal{T}|} y_{\mathcal{T}_j} \cdot f(\mathcal{T}_j) \right) , \\ & y^*_{\mathcal{T}} \geq 0 \ . \end{aligned} \tag{3.11}$$

We search the path frequency distribution whose edge-vise aggregates minimize the difference to the empirical data at the measurement locations. This process is subject to the solver and is schematically depicted in Figure 3.5 [Liebig *et al.* 2010]. As solver, we apply the quasi-Newtonian method L-BFGS-B [Zhu *et al.* 1997] which efficiently estimates the solution under given constraints using a gradient based method. The algorithm is available in the programming language R [R Development Core Team 2009] which we chose as implementation programming language.

The second term of the objective function provides an additional term, which represents the movement pattern heuristic. Reasonable movement pattern heuristics would be for example to prefer the routes with smaller detours [Li *et al.* 2008]. In this case, a detour $f(\mathcal{T}_j)$ can be calculated as the ratio between the path length of $\mathcal{T}_j$ and the minimal path length with the same start and end segment. The term does not just penalize detours, but has the advantage to make a specific solution more invariant on the used solver for the program, because it further restricts the set of possible solutions.

The number of people per valid path that traverses the train station, computed in the previous step, is (according to equation 3.8) used to estimate pedestrian quantities for all locations (i.e. edges) in the closed environment covered by valid paths which contain at least one measurement.

However, the algorithm itself is not sufficient to answer the question from Chapter 1 in an industrial use case scenario. One step towards an industrial applicable system is the software integration of the hereby described method. This software integration is discussed in [Liebig & Xu 2012] and as well focus of Chapter 5. We apply the presented method to an industrial scenario focussing on major Swiss train stations in Section 6.2. Next, the properties, *robustness* and *complexity*, of this method are discussed.

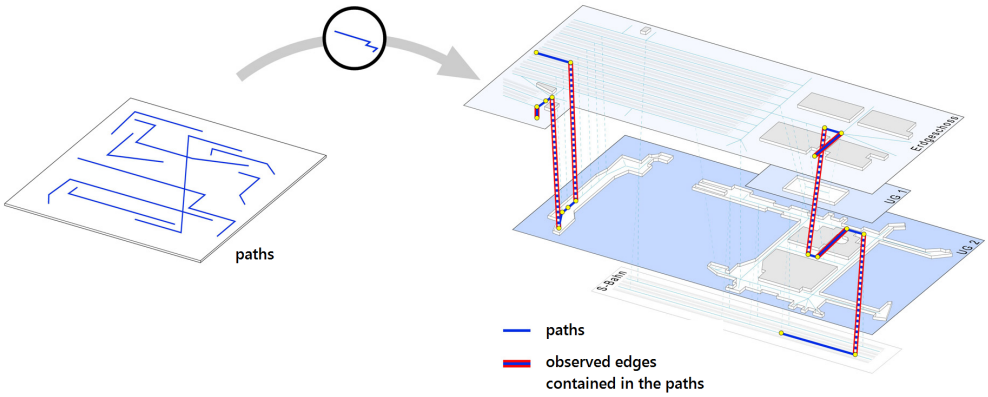

Figure 3.5: Route Based Regression Workflow (image credits [Swiss Poster Research Plus 2010]).

3.6.1 Robustness

Due to the nature of empirical measurements, the pedestrian quantity measurements include small deviations, i.e. erroneous measurements. Thus, the combination of these counts in target quantities (i.e. scope of the program described above, Equation 3.11) lead to *contradictions*. In case the un-preprocessed (original) measurements distinguish movement directions (provided by e.g. manual pedestrian flow counting, see Section 2.3.3), the contradictions can be easily recognized in the raw data by violations of Kirchhoff's law at junctions, whereby the number of incoming people has to equal the number of the outgoing ones. This law was postulated in [Kirchhoff 1845] for electrical circuits. In [Ford & Fulkerson 1962] Kirchoff's law is generalized to arbitrary network flows. In general, the problem of inconsistent pedestrian quantities may arise along multiple junctions. In this case, and in case of undirected quantities, it is harder to pre-identify the contradictions. Therefore, we require the frequency estimation algorithm to recognize such cases automatically and to eliminate them from the model, if required.

Our approach fulfils all of these criteria. The Kirchhoff law holds automatically, because the enumerated path set does: every single path holds the constraint at any junction that the number of incoming people equals to the number of outgoing ones. Multiplying with the path quantities, our algorithm increases the number of people on the paths, but nevertheless the equilibrium [Kirchhoff 1845] remains fulfilled. If each single route fulfils this constraint, the set of all paths including their final quantities also does. Therefore, in our pedestrian model Kirchhoff's law holds.

From this property follows directly an invariance of the method on graph homeomorphisms for the traffic network. This means, if an edge becomes divided into two connected parts, the resulting quantities remain the same. Kirchhoff's law ensures that the quantities of the divided edges equal the undivided ones. Furthermore, small perturbances included in the countings are corrected by the Least Squares Re-

gression. Without the need of a pre-analysis of outliers (caused by deviations of the pedestrian quantity measurement technology) the pedestrian quantity output is defined such that the differences at the measurement locations are minimal, compare Equation 3.11.

3.6.2 Complexity

One weakness of the existing microscopic models discussed in Section 2.4.2 is their poor adjustability by empirical traffic observations [Kretz & Schreckenberg 2006b]. The microscopic models are hard to use, as they require a repeated execution of the whole simulation and an adjustment of the parameters [Kretz & Schreckenberg 2006b]. The time-complexity of each simulation run depends on the number of pedestrians and the length of the explored time-interval. By contrast, the presented approach scales well and does not depend on number of pedestrians nor the chosen time-interval, but the size of the area under observation and the number of sensors, as the solver L-BFGS has a linear-time iteration complexity [Schmidt *et al.* 2011]. This property is important in the industrial case, where we made calculations for the 27 major train stations in Switzerland, including some with a daily usage of several hundred thousand passengers per day (see Chapter 6).

3.7 Approach with Pattern Knowledge: GPR

In this section we focus on the pedestrian quantity estimation in closed environments, e.g., train stations, shopping malls and zoos. Unlike the outdoor pedestrian 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 [Andrienko *et al.* 2012] 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 time-stamp t. The location based aggregate, undirected flow counts, for time interval Δt, also known as quantity or traffic frequency, is defined as (compare Definition 8) $NM_u(e, \Delta t)$.

Here, we propose a nonparametric method[6], Gaussian Process (GP) with a random-walk based trajectory kernel. The method explores not only the commonly used information in the literature, e.g. traffic network structures (retrieved by tessellation of the floor plan sketch) and recorded (or aggregated) undirected flow counts NM_u at measurement locations, but also the move preferences of pedestrians (movement patterns) retrieved from sensors or the local experts.

[6]**published with a major contribution of the author in:**
T. Liebig, Z. Xu, M. May and S. Wrobel. *Pedestrian Quantity Estimation with Trajectory Patterns*. In Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases ECML PKDD 2012, Part II, LNCS 7524, pages 629–643. Springer, 2012
T. Liebig, Z. Xu and M. May. *Incorporating Mobility Patterns in Pedestrian Quantity Estimation and Sensor Placement*. In J. Nin and D. Villatoro, editors, Proceedings of the First International Workshop on Citizen Sensor Networks CitiSens 2012, LNAI 7685, pages 67–80. Springer, 2013

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 $\mathbf{y} = \{y_s := NM_u(\tilde{e}_s, \Delta t) : s = 1, \ldots, S\}$. Additionally, we have the information of the (major) pedestrian movement patterns $\mathcal{T} = \{T_1 \rightarrow T_2 \rightarrow \ldots\}$ within the traffic network, collected from the local experts or the tracking technology (in our applications we use Bluetooth tracking technology). Obviously, taking into account the movement patterns is beneficial to predict the unknown pedestrian quantities: the edges included in a movement pattern appear to have similar pedestrian quantities. To overcome 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 [De Raedt 2008, Getoor & Taskar 2007], commonly used GP relational methods are to introduce an unobserved, i.e., *latent variable* (also known as *hidden variable*) to each vertex, and the values of edges is therefore modeled as a function of latent variables of the involved vertices, e.g. [Yu *et al.* 2006, Chu *et al.* 2006]. Although these methods have the advantage that the problem complexity remains linear with increasing number of 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 [Gong & Wang 2002, Lam *et al.* 2006, Neumann *et al.* 2009] within a Gaussian process regression framework. Firstly, we convert the original network $\tilde{\mathcal{G}}(\tilde{\mathbf{V}}, \tilde{\mathbf{E}})$ to a line graph [Harary & Norman 1960] $\mathcal{G}(\mathbf{V}, \mathbf{E})$ that represents the adjacencies between edges of $\tilde{\mathcal{G}}$.

Definition 17 (Line Graph) *In a line graph $\mathcal{G}(\mathbf{V}, \mathbf{E})$ of a graph $\tilde{\mathcal{G}}(\tilde{\mathbf{V}}, \tilde{\mathbf{E}})$, each vertex $v_i \in \mathbf{V}$ represents 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 $\tilde{\mathcal{G}}$ [Harary & Norman 1960].*

To each vertex $v_i \in \mathbf{V}$ in the line graph, we introduce a latent variable f_i which represents the true pedestrian quantity at v_i. It is the value of a function over the line graph and the known movement patterns, as well as the possible information about the features of the vertex (e.g. spatial attributes). The observed pedestrian quantities (within a time interval Δt) are conditioned on the latent function values with Gaussian noise ε_i

$$y_i = f_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \ . \tag{3.12}$$

Since the parameters of the function y_i 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 $k(x_i, x_j)$ [Rasmussen & Williams 2006]. 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 [Rasmussen & Williams 2006]

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')) \ . \tag{3.13}$$

Without loss of generality, we assume the mean equals zero so that the GP is completely specified by the covariance function. Formally, the multivariate Gaussian prior distribution of the function values **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 vertex features $\mathbf{x} = \{x_1, ..., x_M\}$ are 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:

$$k(x_i, x_j) = \kappa^2 \exp\left(-\frac{\rho^2}{2}\sum_{d}^{D}(x_{i,d} - x_{j,d})^2\right), \tag{3.14}$$

where κ and ρ are hyperparameters.

Since the latent variables **f** are linked together into a line 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:

$$K = \left[\beta(L + I/\alpha^2)\right]^{-1}, \tag{3.15}$$

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 [Neumann *et al.* 2009], 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 (movement patterns, collected from the local experts or tracking technology) into the model, we explore a graph kernel inspired from the diffusion process [Kondor & Lafferty 2002].

Assuming that a pedestrian randomly moves on the line 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 a diffusion process, the similarity between vertices v_i and v_j is defined as

$$s(v_i, v_j) = \left[\sum_{k=1}^{\infty}\frac{\lambda^k}{k!}A^k\right]_{ij}, \tag{3.16}$$

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 [Kondor & Lafferty 2002]:

$$K = \left[\sum_{k=1}^{\infty} \frac{\lambda^k}{k!} L^k\right] = \exp(\lambda L) \ . \tag{3.17}$$

On a traffic network within closed environment, the pedestrian will move not randomly, but with respect to a set of movement patterns. In case of the line graph, they are denoted as sequences of vertices, e.g.,

$$\left\{ \begin{array}{ll} T_1 & = v_1 \to v_3 \to v_5 \to v_6, \\ T_2 & = v_2 \to v_3 \to v_4, \\ T_3 & = v_4 \to v_5 \to v_1, \\ & \ldots \end{array} \right\} . \tag{3.18}$$

Each movement pattern T_ℓ can also be represented as an adjacency matrix in which $\hat{A}_{i,j} = 1$ iff $v_i \to v_j \in T_\ell$ or $v_i \leftarrow v_j \in T_\ell$. The patterns are subsequences of the trajectories. For example, the patterns of T_1 are $\{v_1 \to v_3, v_3 \to v_5, v_5 \to v_6, v_1 \to v_3 \to v_5, v_3 \to v_5 \to v_6\}$. Given a set of movement 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 patterns of $\mathcal{T}$. Thus we have

$$\begin{aligned} \hat{s}(v_i, v_j) &= \left[\sum_{k=1}^{\infty} \frac{\lambda^k}{k!} \hat{A}^k\right]_{ij}, & \hat{K} &= \left[\sum_{k=1}^{\infty} \frac{\lambda^k}{k!} \hat{L}^k\right] = \exp(\lambda \hat{L}) \\ \hat{A} &= \sum_{\ell} \hat{A}_\ell, & \hat{L} &= \hat{D} - \hat{A}, \end{aligned} \tag{3.19}$$

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 $(\mathbf{y}, \mathbf{f}_u)^T$ follow a Gaussian distribution

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_u \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \hat{K}_{\overline{u},\overline{u}} + \sigma^2 I & \hat{K}_{\overline{u},u} \\ \hat{K}_{u,\overline{u}} & \hat{K}_{u,u} \end{bmatrix}\right), \tag{3.20}$$

where $\hat{K}_{u,\overline{u}}$ are 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:

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

This Gaussian distribution denotes the mean of the unobserved traffic frequencies m

and their covariances Σ.

3.7.1 Sensor Placement

Besides pedestrian quantity estimation, incorporating movement patterns also enables effectively finding sensor placements that are most informative for pedestrian quantity estimation on the whole traffic network. Thus, for sensor placement we identify the most informative locations. This is utilized by the GP model of the pedestrian quantities [Krause *et al.* 2006].

To identify the most informative locations $\mathcal{I}$, we employ the exploration strategy, maximizing mutual information [Krause *et al.* 2006]

$$\arg\max_{\mathcal{I}\subset\mathbf{V}} H(\mathbf{V}\backslash\mathcal{I}) - H(\mathbf{V}\backslash\mathcal{I} \mid \mathcal{I}) \ . \tag{3.21}$$

This is equivalent to finding a set of vertices $\mathcal{I}$ in the line graph that 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 and does not involve any pedestrian quantity observations. To solve the optimization problem, we employ a polynomial time approximate method [Krause *et al.* 2006]. In particular, starting from an empty set $\mathcal{I} = \emptyset$, each vertex is selected with the criterion:

$$v_* \leftarrow \arg\max_{v\in\mathbf{V}\backslash\mathcal{I}} \ H_\varepsilon(v \mid \mathcal{I}) - H_\varepsilon(v \mid \overline{\mathcal{I}}) \ , \tag{3.22}$$

where $\overline{\mathcal{I}}$ denotes the vertex set $\mathbf{V}\backslash(\mathcal{I}\cup v)$. $H_\varepsilon(x|Z) := H(x|Z')$ 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 $H_\varepsilon(x|Z)$ is computed as

$$\begin{aligned} H_\varepsilon(x \mid Z) &= \frac{1}{2}\ln 2\pi e\sigma^2_{x|Z'} \\ \sigma^2_{x|Z'} &= \hat{K}_{x,x} - \hat{K}^T_{x,Z'}\hat{K}^{-1}_{Z',Z'}\hat{K}_{x,Z'} \ . \end{aligned} \tag{3.23}$$

The term $\hat{K}_{x,Z'}$ are 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 movement pattern kernel, the pedestrian quantity observations at the vertices selected with the criterion (3.22) can well estimate the situation of the whole traffic network (compare validation in Section 3.7.2 and Chapter 6). Chapter 6 shows a successful application to the zoo of Duisburg.

3.7.2 Validation

We compare our Gaussian Process Regression method with state-of-the-art traffic volume estimation approaches (compare Section 3.4).

In the experiments we use the Spatial S-kNN method [May *et al.* 2008] with

distance-weighted 5 nearest neighbours (as [Neumann *et al.* 2009] does). Particularly this method detects for each unmeasured edge the 5 nearest neighbours among the measured ones. Their traffic frequencies are weighted by distance to achieve prediction for the unmeasured ones. Since the Spatial kNN is a geometric algorithm which requires spatial representation of the traffic network, we apply the Fruchterman Reingold algorithm [Fruchterman & Reingold 1991] to lay out the synthetic 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), introduced in previous Section 3.7. The prediction performance of the methods is measured with mean absolute error (MAE) $MAE = n^{-1} \sum_{i=1}^{n} |y_i - f_i|$.

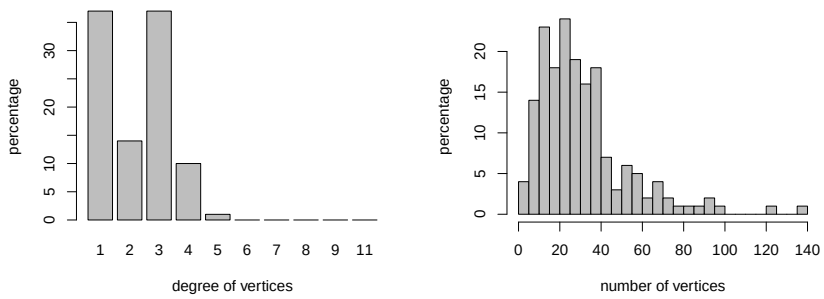

Figure 3.6: Distributions of vertex-degree (left) and number of vertices (right) among 170 large German train stations [Liebig *et al.* 2012b].

To approximate the true situation, we study traffic networks of the 170 largest public train stations in Germany. The distributions of vertex-degree and vertex-number are visualized in Figure 3.6. 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 [Viger & Latapy 2005] 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 remain permanently 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. The validation is not performed on the original 170 traffic networks as they are a confidential part of an industrial project.

Experimental results are depicted in Figure 3.7. 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 box plots (see Appendix B

for an introduction to box plots). Figure 3.7 provides a condensed view, as every single box-whisker plot in the diagram (compare Appendix B) represents 100 runs of the algorithm. The corresponding condensed result tables are in Appendix C. Throughout the tests, our method achieved minimal MAE and minimal median MAE, and therefore best results for the pedestrian quantity estimation problem. The proposed method outperformed commonly used S-kNN approach, especially when traffic networks are sparsely monitored. With increasing 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 movement pattern kernel achieved best performance compared to other kernels.

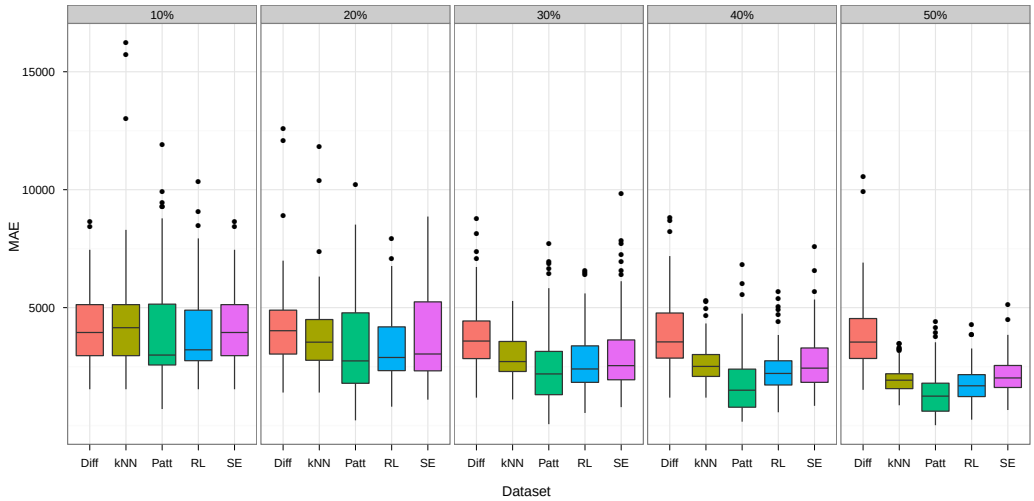

Figure 3.7: Pedestrian quantity estimation on networks of train stations. Performance is measured by MAE at settings with different ratios of monitored edges (10 to 50 percent from left to right). The five methods: GPR with diffusion kernel (Diff), spatial k-nearest neighbour (kNN), GPR with trajectory pattern kernel (Patt), GPR with regularized Laplacian (RL) and GPR with squared exponential kernel (SE) [Liebig *et al.* 2012b].

We validate the performance of the sensor placement in the visitor monitoring application scenario, in Section 6.3.6. In brief, the sensor placement we proposed, outperforms random placement in terms of estimation error. Furthermore the method is also promising if the initial presumption on closed environments (Kirchhoff's law is fulfilled for pedestrian quantity within considered time interval) is violated, we tested applicability for sensor number reduction (shown for a soccer stadium dataset in Section 6.4.4).

3.8 Summary

Estimation of pedestrian quantities is an important question to various applications. Utilizing the preliminary fundamentals, this chapter focused on pedestrian quan-

tity estimation using movement patterns. We presented two complementary regression approaches Least Squares Regression (LSR) and Gaussian Process Regression (GPR) to tackle the task with different input data and constraints on movement. Though existing approaches do not incorporate naturally occurring movement patterns, we showed that these patterns contain valuable information on spatial correlations among traffic quantities.

In previous sections the presented methods are tested and analysed individually. One common advantage of the presented methods is the independence of the run time from the number of modelled pedestrians. Contradictions may occur in manual pedestrian quantity measurements (presented in Section 2.3.3). The LSR method can identify these contradictions occurring in empirical sensor readings, which are likely in case of short term traffic counts [FDOT 2012], discussed in Section 3.6.1. The GPR method not just outperforms existing ADT estimation methods, but supports placement of a constrained number of sensors among the site.

In order to stress the benefits of our contributed methods, next, we revive the introductory example at the T-Junction network consisting of only one junction, Section 3.1. As shown in Figure 3.1, 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 [Liebig *et al.* 2012b]. In Section 3.5, we applied existing state-of-the-art quantity estimation approaches, and discussed why they fail in modelling the pedestrian quantities correctly. Here, we apply the novel algorithms LSR and GPR to the problem, and achieve good results, as the knowledge on movement patterns is incorporated by the two methods, see Figure 3.8 and compare Sections 3.6 and 3.7 for details on the proposed methods. In Figure 3.8 the ground truth is depicted, with a relative frequency of 100% in the horizontal corridor. Given the left handed frequency to the estimation algorithm results by application of the LSR method in a correct traffic estimation (Figure 3.8 - middle) and in case of the GPR for an almost correct estimation (Figure 3.8 - right), with a relative estimation error at the unobserved corridors of 0.5%. These results outperform related methods, compare Figure 3.3.

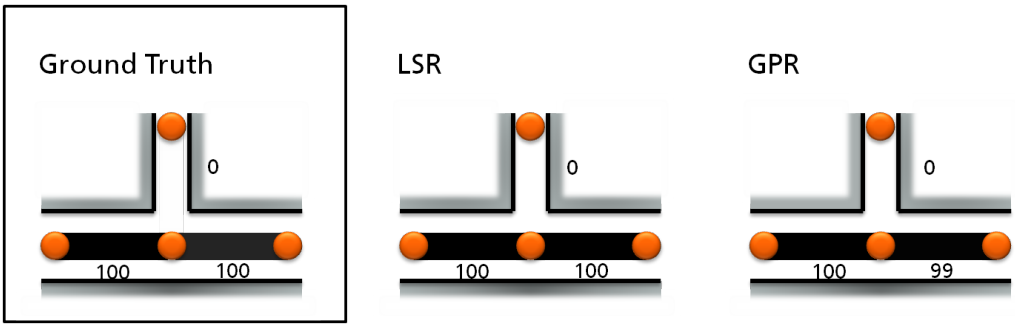

Figure 3.8: Application of our proposed methods to the T-Junction example.

Whereas this chapter focussed on algorithmic aspects of the quantity estimation and validated the presented approaches, following chapters address the pattern anal-

ysis from observation data and describe a software system for pedestrian mobility analysis as well as successful industrial application of the presented methods.

Chapter 4

Movement Pattern Analysis based on Bluetooth Tracking Data

"It is a capital mistake to theorize before one has data."

—Sir Arthur Conan Doyle[1]

Contents

Previous chapters described the natural occurrence of mobility patterns in pedestrian mobility. We motivated and discussed the crucial task of pedestrian quantity estimation which gives indispensable input to location evaluation, attractor identification or event monitoring scenarios. Two novel algorithms were presented that incorporate these patterns (either direct as sequences of visited places or via a heuristic) to achieve estimations for the pedestrian quantities.

In this chapter, we describe, how to acquire these patterns from observation data. Thus, we utilize recently evolved Bluetooth tracking technology which records *Episodic Movement Data* from pedestrian mobility and describe methods for analysis of this particular type of data.

[1]Scottish physician and writer, 1859–1930, A Study in Scarlet [Doyle 1969]

Next chapters present a fusion of the quantity estimation and pattern analysis methods with sophisticated state-of-the-art sensor technologies to a system for pedestrian mobility analysis. Real world applications are presented afterwards.

4.1 Introduction

In Chapter 2 we presented a comprehensive overview on state-of-the-art pedestrian tracking technologies. We started this survey with video surveillance and 3D laser scan data continued by radio frequency based tracking technologies, e.g. GSM. In this context, recently evolved Bluetooth scanners were also introduced. Next, we present the latest contributions on Bluetooth tracking by the author of this thesis. Thus, we study (after a recall of Bluetooth tracking technology) (1) the representativeness of the recorded data and (2) the possibility to retrieve knowledge on microscopic movement [Liebig & Kemloh Wagoum 2012, Utsch & Liebig 2012] as well as (3) macroscopic analyses (spatial dependency models and temporal similarity clustering). *Episodic Movement Data* is recorded by Bluetooth tracking since it holds the common uncertainties on *Accuracy*, *Coverage* and *Continuity*. After a brief introduction to episodic movement data, we present our contributions on the (visual) analysis of movement patterns and their probabilistic modelling.

Thus, we show that Bluetooth tracking data naturally preserves the sequence patterns of pedestrian movement and therefore can be handed in directly to the Gaussian Process pedestrian quantity estimation algorithm which was subject of the previous chapter. This procedure (handing in the Bluetooth data which contains sequence patterns directly to the quantity estimation algorithm) will be applied in Chapter 6 to different application scenarios.

4.2 Bluetooth Tracking

In GPS-less environments novel technologies are required for pedestrian mobility recording. Traditional approaches are surveys and video surveillance which are discussed in Section 2.3. Both methods are relatively difficult to realize. Surveys need to draw a representative sample of the pedestrians both in space and time. In turn, video surveillance systems require additional scaffoldings that need to be integrated in existing architectures. Furthermore, their accuracy depends on weather and light conditions.

The need for further robust passive localization technologies pushed the development of sensors that are capable to monitor people's movement. First choice is to track most popular digital gadgets: mobile phones and intercoms. Analysis of mobile network GSM (Global System for Mobile Communications) log files causes strong privacy objections [Giannotti & Pedreschi 2008]. This problem could be tackled by processing the mobility data locally in the device. Recent work utilizes mobile devices to monitor location based events (*visits* [Kopp *et al.* 2012], *moves* [Hoh *et al.* 2012]) or even more complex *movement patterns* [Florescu *et al.* 2012]. However, so far their work just processes streams of GPS position updates and efforts are required in order to utilize

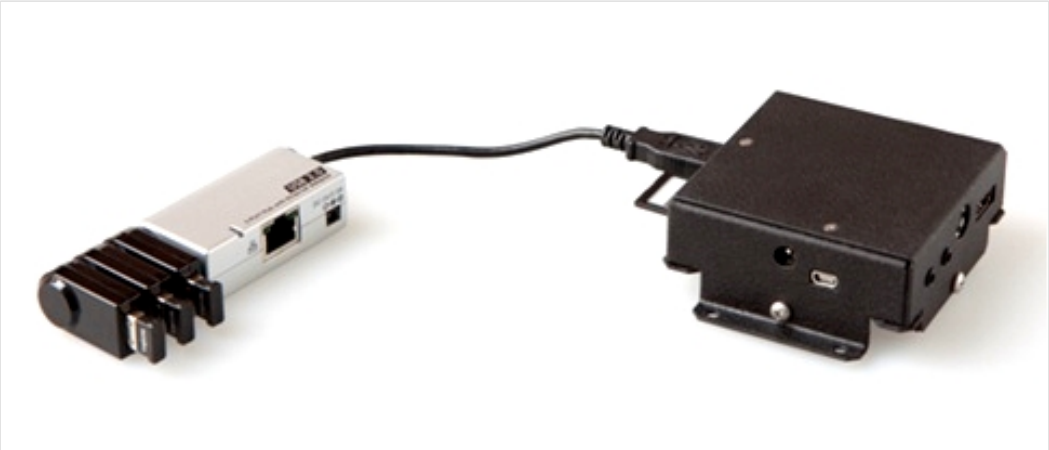

Figure 4.1: Bluetooth Scanner developed at Fraunhofer IAIS.

different radio frequency based positioning technologies, moreover, besides privacy issues, the GPS and GSM data does not support precise indoor positioning.

Besides these existing approaches, Bluetooth tracking technology is an emerging technology for combined indoor/outdoor monitoring tasks [Fuller 2009, Bruno & Delmastro 2003, Kolodziej & Hjelm 2006]. Bluetooth tracking belongs to the radio frequency based tracking technologies. In the last years many of these radio frequency technologies (GSM, GPS, Bluetooth, etc.) have emerged, each of them providing different positioning techniques. These positioning techniques bare different advantages and disadvantages for various applications. They divide into following five classes [Fuller 2009]:

- **Proximity** is a cell based technology. Whenever a tracked device is close enough to a sensor to establish radio frequency communication (this is similar to accessing the sensors footprint) the location of the tracked device is mapped to this particular cell. Usage of multiple sensors offers the possibility to track movement.

- **Direction Finding (DF)** utilizes rotatable sensor areas and utilize the proximity approach for a rotating sensor. Whenever a tracked device comes into the sensor area, the direction of its finding is known and its position can be estimated by use of multiple sensors.

- **Angle Of Arrival (AOA)** computes the angle between the sensor and the device. The precision of (AOA) decreases with increasing distance due to the increasing influence of dispersion of the radio waves and their echo. AOA requires a free line of sight between sensor and tracked device. The utilization of two or more sensors allows positioning of tracked device by angulation methods (compare for example [Picard & de La Hire 1780] for an introduction to triangulation).

- **Timing and Phase Of Arrival (TOA, POA)** computes the time difference the signal needs to traverse the distance between sensor and tracked device. In case of multiple sensors, the device can compute distances to these sensors using a

trilateration algorithm [Torge 2001]. Larger distances improve precision of this localization technology. Time Of Arrival requires a precise synchronisation of the involved devices.

- **Radio Signal Strength (RSS)** utilizes the dependency among signal strength and distance. In theory, there exists an inverse proportional relation among these two values, i.e., the higher the distance, the lower the signal strength. RSS localization is easily to realise, i.e. no additional processing of received values is required (as it is for AOA and TOA). Similar to TOA, the position can be computed by trilateration algorithms.
- **Doppler** uses both signal strengths from the tracked device to the sensor as well as from the way back from the sensor to the tracked device. The signal, which is sent from the tracked device, is compared to the one sent by the sensor. Thus, the disturbances on the signal strengths have a lower impact on the sensor values. The technology requires access to the data at the sensor and at the tracked device.

So far recently evolved Bluetooth Tracking sensors utilize various positioning techniques from the ones mentioned above: Proximity, RSS and DF. Thus, such sensors have been used in multiple scenarios: event monitoring at a soccer match in France [Liebig & Kemloh Wagoum 2012] and a car race [Stange *et al.* 2011] with the author's contribution. There, a mesh of Bluetooth sensors has been placed at carefully selected indoor as well as outdoor locations. Besides event monitoring, also other successful indoor applications of Bluetooth scanners are described in literature. In [Pels *et al.* 2005] various scanners were placed at Dutch train stations to record transit travelers. Accurate locating and following of objects within complex facilities is as well an important research topic [Hallberg *et al.* 2003]. So far Bluetooth tracking is mostly used to monitor a sample of visitors [Liebig & Kemloh Wagoum 2012, Andrienko *et al.* 2012, Hallberg *et al.* 2003] and extract their route choices [Liebig & Kemloh Wagoum 2012, Utsch & Liebig 2012]. The work presented in [Hagemann & Weinzerl 2008] uses Bluetooth tracking to track people among a public transportation network, whereas [Leitinger *et al.* 2010] gives a general overview on possibilities using Bluetooth tracking technology. The work in [Liebig & Kemloh Wagoum 2012] extracts the route choices of the visitors and hands them to an agent-based pedestrian microsimulation in order to extract microscopic movement values. Whereas [Utsch & Liebig 2012] uses the radio signal strength value directly for localization and route choice detection in complex environments.

4.2.1 Sensor Technology

The Bluetooth sensor devices (also scanners, transceivers or beacons) throughout this thesis are inspired by [Bruno & Delmastro 2003] and are assembled using a microcomputer and three USB Bluetooth antennas. A Linux based software activates the antennas inquiry mode and logs the (hashed [National Institute of Standards and Technology 2002]) MAC addresses of the detected devices[2]. The scan interval of approximately $10.24s$ [Woodings *et al.* 2001, Bruno & Delmastro 2003] is (theoretically)

[2] A similar software for Bluetooth Tracking is published by the University of Ghent at https://github.com/Rulus/Gyrid with GPL licence, last accessed 06/30/2012

reduced to its third by utilization of three antennas which are started with delays. This theoretic time span is required for the inquiry process consisting of frequency hopping and device discovery [Bluetooth SIG 2004]. However, in practice the three antennas are not synchronous and inquiry scans are not performed with a constant frequency. This challenge was already addressed in the section on *episodic movement data* (Section 2.2.4). The recorded data log entries consists of:

```
[time stamp], [sensor ID], [sha256(MAC)], [signal strength] .
```

The Bluetooth antennas are available in different types and the ones that we assembled have a range of $20m$ or $100m$. Nevertheless, since the inquiry mode requires communication in both directions (from the sensor to the mobile device and vice versa) [Bluetooth SIG 2004], the range of the sensors does not only depend on the sensor's antennas but also on the antenna in the mobile device (and its configuration). Thus, lower sensor ranges of about $20m$ are expected and verified in preliminary tests. The recorded unique Media Access Control addresses (MAC) of the Bluetooth antennas of the mobile devices consists of six bytes [IEEE 2002]. Three of them depend on the vendor [Bluetooth SIG 2004, IEEE 2002] and provide valuable information for analysis.

In [Andrienko *et al.* 2012] and previous section, Section 2.2.4, the formalization of recorded data is addressed. There, similarities to other episodic movement data are presented as well as methods for their aggregation and visual representation.

However, since Bluetooth tracking just monitors a sub-sample of the population, it is also important to study its spatio-temporal representativeness. One requirement for representativeness of the recorded sub-sample is a constant ratio of detected devices (to the total number of people) in space and time. This issue is addressed in the following section at a soccer stadium.

4.2.2 Representativeness Analysis

Bluetooth tracking monitors just a sub-sample of the pedestrians and these pedestrians walk in groups (see Section 2.1.3) and have route preferences and thus are not distributed uniformly at random in space and time. Expected representativeness is about 10% [Versichele *et al.* 2012a]. In order to analyse whether this sub-sample is representative for all the pedestrians, additional analysis are required.

Thus, we place in a multi-purpose arena with digital access control Bluetooth scanners next to the entry gates, see map depicted in Figure 4.2. The data was collected during a soccer match on 15/05/2012 with approximately 43,000 visitors. In order to compare the access control data with the Bluetooth measurements, it was not necessary to cover all gates by scanners, since the access control provides detailed visitor numbers for every single gate. As seen in the map, we covered the two main entrances (north-west and south) to the arena. During the soccer match 36,734 persons passed the two gates in total (this is the ground truth recorded by the access control). In turn, Bluetooth tracking recorded 2,581 which results in a detection ratio of 7.03%.

Given this ratio, further investigation is required to ensure whether the recordings of the Bluetooth scanners are representative for all visitors in space and time. In order

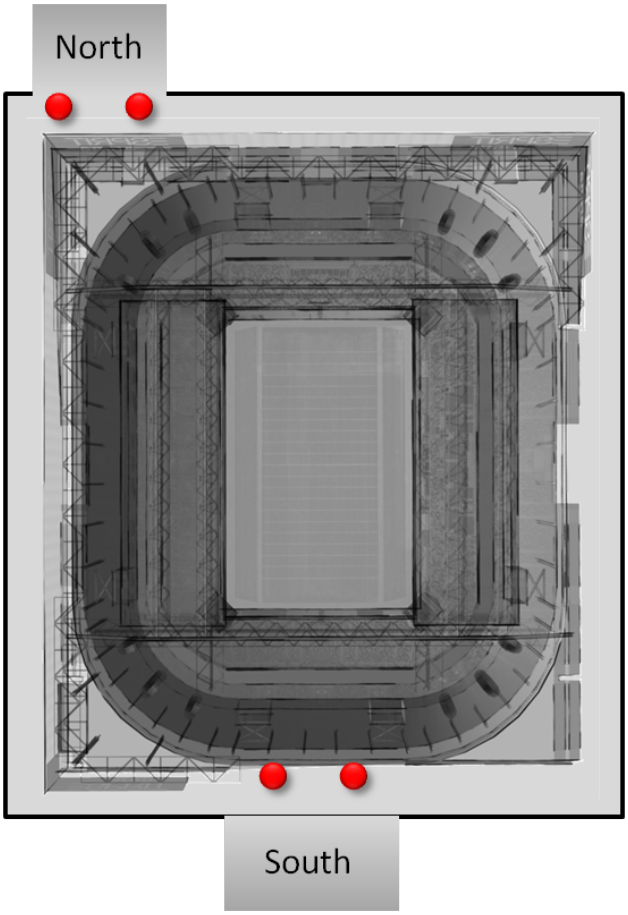

Figure 4.2: Locations of the Bluetooth scanners (red dots) at the multi-purpose arena.

to analyse representativeness it is necessary to consider both extents of the spatio-temporal data. Next we analyse whether the detection ratio persists among the two gates at every time interval.

Bluetooth tracking records a person during its stay in a scanner footprint. Thus, multiple data points can be available for a particular person in a sensor log file. As the scanners are located at the gates, the first and the last data point for a person can be considered to represent its *entry* and *exit*. However, because the access control generated just one data point for every single person at its arrival, the comparisons are performed with the entry-event, which results from the first log-entry for a particular Bluetooth enabled device.

First, we evaluate the spatial representativeness. Therefore, we analyse which entry the visitors have used to access the arena. The access control recorded 66% of the (considered 36.734) people at the southern gate. In comparison, Bluetooth recorded 65% (of the recorded 2.581) at the southern gate. Hence, the spatial distribution of the people is similarly spatially distributed.

Next, the temporal representativeness is analysed. The entry events are aggregated in 30 minute intervals and normalized by their total number. This results in

a percentage of the recorded visitors for every time slice. The values are plotted together in Figure 4.3. Furthermore, the computation of the correlation among the two time graphs returns 0.997. The same analysis is repeated with shorter time intervals of just 15 minutes. These analyses are as well shown in Figure 4.3. The correlations in these cases are slightly lower: returning 0.982 at the southern gate and 0.987 at the northern gate. These high correlations justify that the data is also time-representative.

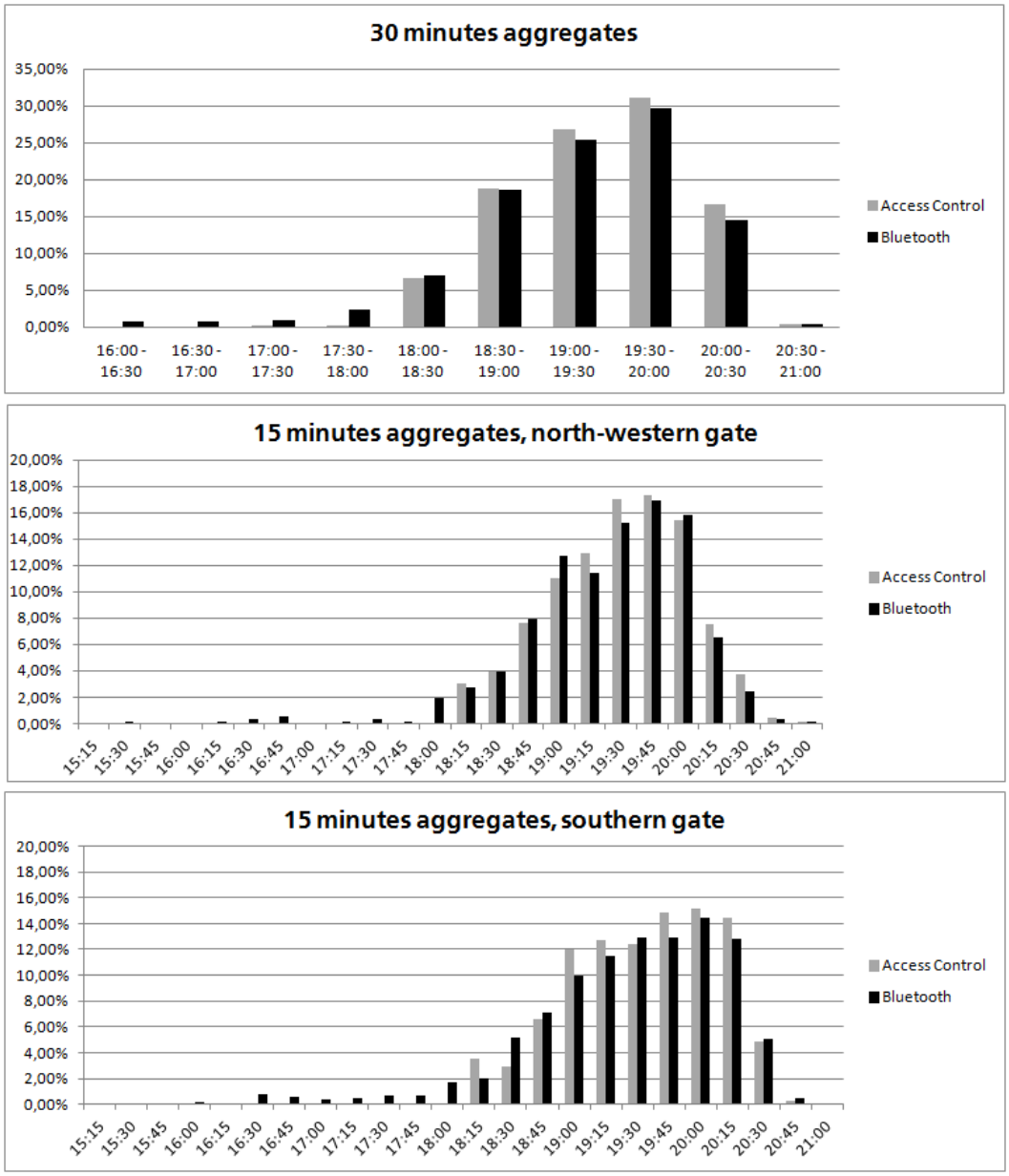

Figure 4.3: Temporal Bluetooth counting representativity in comparison to access control.

In summary, the presented empirical analysis confirmed spatio-temporal representativeness and we obtained very high correlation and accuracy in both time and space dimensions. Thus, the Bluetooth tracking data can be scaled in order to retrieve the total visitor number.

4.3 Microscopic Movement Analysis using Bluetooth

Besides the total number of persons at the sensor locations and the flows among the sensors, the individual position and route choice might be of interest as it provides the fundamentals for intelligent location aware environments. Such detailed movement information is also called microscopic movement. In this section, we briefly mention two complementary approaches to derive microscopic movement data from Bluetooth tracking. First one applies agent based simulation to reconstruct people's movement [Liebig & Kemloh Wagoum 2012] and the other one makes use of the radio signal strength to reconstruct individual positions and path choices [Utsch & Liebig 2012]. This is not the main focus of the thesis, but it highlights the capabilities of Bluetooth tracking and provides a comprehensive overview on the applicabilities in combination with the next section on macroscopic movement analysis.

4.3.1 Modelling Microscopic Movement using Micro-Simulation

In the joint work [Liebig & Kemloh Wagoum 2012] we propose the combination of an agent-based pedestrian simulation (the Generalized Centrifugal Force Model GCFM [Chraibi *et al.* 2011], compare Section 2.4.2.1) with Bluetooth tracking data. The GCFM utilizes navigation graphs to determine intermediate goals of the agents. This could be easily matched with the Bluetooth sensor positions. Therefore, the pedestrian simulation represents the realistic route choice distribution. We refer to [Liebig & Kemloh Wagoum 2012] for more details.

4.3.2 Monitoring Microscopic Movement based on Bluetooth Radio Signal Strength

Another possibility for reconstruction of the individual position and route choice was proposed in [Utsch 2011]. We extended this approach in the joint publication [Utsch & Liebig 2012]. Bluetooth sensors were placed equidistant with overlapping footprints. The radio signal strengths of multiple sensors are combined to a footprint. These footprints are compared to reference measurements. This process can reveal individual positions (and route choices as well) based on Bluetooth tracking. For more details, we refer to [Utsch & Liebig 2012].

4.4 Macroscopic Movement Analysis using Bluetooth

The popularity of cellular phones and advances in information and sensor technologies lead the way towards new location recording techniques and thus new types

of movement data. Whereas previous section addressed the investigation of individual mobility, this section tackles the macroscopic movement models. Based on the episodic sensor readings derived from Bluetooth tracking technology, we create spatio-temporal aggregates of people's presence and flow which are subject for analysis.

Episodic Movement Data (Section 2.2.4) refers to data about spatial positions of moving objects where the time intervals between the measurements may be quite large and therefore the intermediate positions cannot be reliably reconstructed by means of interpolation, map matching, or other methods[3]. Such data can also be called *temporally sparse*; however, this term is not very accurate since the temporal resolution of the data may greatly vary and occasionally be quite fine. There are multiple ways of data collection producing episodic movement data:

- **Location based:** Positions of objects are recorded only when they come into the range of static sensors. The temporal resolution of the collected data depends on the coverage and density of the spatial distribution of the sensors.
- **Activity based:** Positions of objects are recorded only at the times when they perform certain activities, for example, call by mobile phones, pay by credit cards or send posts to a community website.
- **Device based:** Positions are measured and recorded by mobile devices attached to the objects but this cannot be done sufficiently frequently, for example, due to the limited battery lives of the devices i.e. when tracking movements of wild animals.

Irrespective of the collection method we can identify three types of uncertainty (depicted in Figure 4.4) [Andrienko *et al.* 2012].

- First, the common type of uncertainty in any episodic movement data is the lack of information about the spatial positions of the objects between the recorded positions (continuity), which is caused by large time intervals between the recordings and by missed recordings.
- Second, a frequently occurring type of uncertainty is imprecision of the recorded positions (accuracy). Thus, a sensor may detect an object within its range but may not be able to determine the exact coordinates of the objects. For a mobile phone call, the localization precision may be the range of a certain antenna but not an exact point in space. Due to these uncertainties, episodic movement data cannot be represented as continuous trajectories, i.e., lines in the spatio-temporal continuum where known (measured) positions are linked by straight or curved segments.
- Third, the number of recorded objects (coverage) may also be uncertain due to the usage of a service or due to the utilized sensor technology. For example, one

[3] **published with a major contribution of the author in:**
N. Andrienko, G. Andrienko, H. Stange, T. Liebig and D. Hecker. *Visual Analytics for Understanding Spatial Situations from Episodic Movement Data.* KI - Künstliche Intelligenz, pages 241–251, 2012.
T. Liebig, G. Andrienko and N. Andrienko. *Methods of Analysis of Episodic Movement Data.* In Mobile Tartu, pages 24–25, 2012

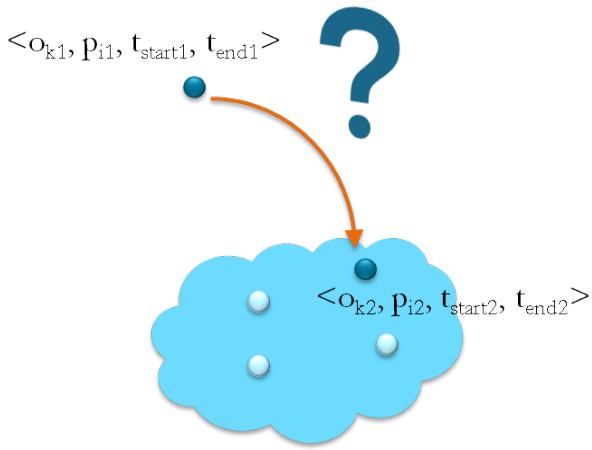

Figure 4.4: Common Uncertainties in Episodic Movement Data, compare Section 2.2.4.

individual may carry two or more devices with Bluetooth transceivers, which will be registered by Bluetooth sensors as independent objects. On the other hand, the sensors only capture devices with activated Bluetooth services. The activation status may change while a device carrier moves from one sensor to another.

Many of the existing visual and data mining methods designed for dealing with movement data are explicitly or implicitly based on the assumption that movement is continuous between the measured positions and are therefore not suitable for episodic data. Interpolation is obviously involved in visual representation of trajectories by continuous lines but it is also implicitly involved in computation of movement speeds, directions, and other attributes characterising the movement (these computations also assume that the positions are precise). The same holds for the summarisation of movement data in the form of density or vector fields. Mining methods for finding patterns of relative or collective movement of two or more objects (e.g. meeting or flocking) also require fine-resolution data. Since many of the existing methods are not applicable to episodic movement data, there is a need in finding suitable approaches for analysing this kind of data. Due to the uncertainties, episodic data are usually not suitable for studying the movement behaviours of individual objects.

In order to overcome these shortcomings, we suggest the aggregation of many individual tracks to compensate for missing data and uncertainties in the spatial and temporal coverage. By example of episodic movement data, we motivate the utilisation of visual and computational methods for analysing complex data. Visual analytics strives at multiplying the analytical power of both human and computer by finding effective ways to combine interactive visual techniques with algorithms for computational data analysis [Keim *et al.* 2008]. The key role of the visual techniques is to enable and promote human understanding of the data. Particularly, visual analytics can help in understanding the data for data mining tasks, such as distributions, features, clusters, patterns. Visual analytics approaches are applied to data and problems for which there are (yet) no purely automatic methods to deal with. By enabling

human understanding, reasoning, and use of prior knowledge and experiences, visual analytics can help the analyst to find suitable ways for data analysis and problem solving, which, possibly, can later be fully or partly automated. Thus, visual analytics can drive the development and adaption of learning and mining algorithms. In the next sections, we describe sequence pattern mining and aggregation of episodic movement data and present the analysis tasks in which the episodic movement data can be used. Then, after discussing the relevant literature, we present our visual analytics methods (clustering and correlation analyses) and tools using an example of episodic movement data collected by Bluetooth sensors.

4.4.1 Sequence Pattern Mining

Movement data naturally contains movement patterns due to the non-random individual movement (compare Section 2.1). In case of episodic movement data, the patterns are still preserved in the data and existing analysis methods can be applied to extract them. In [Kisilevich *et al.* 2010] sequence pattern acquisition on episodic movement data (in this case geo-tagged photos) is performed using the Teiresias algorithm [Rigoutsos & Floratos 1998]. This algorithm was designed for biological sequence mining and it provides three important parameters for pattern search:

- L - number of literals in the pattern,
- K - minimum number of occurrences of the pattern,
- W - minimum length of pattern.

Given these three parameters, the Teiresias algorithm performs a frequent pattern search for *Episodic Movement Data*, which gives insights on movement preferences.

However, the Bluetooth tracking data preserves movement patterns, and no pattern mining is required for handing it in the quantity estimation algorithm. Next sections present more practical methods for visual inspection of spatial or temporal correlations in the recorded data.

4.4.2 Spatio-Temporal Aggregation

Episodic movement data consist of records including the following components: object identifier o_k, spatial position p_i, time t, and, possibly, other attributes. The spatial position may be specified directly by spatial (geographic) coordinates $p = (x, y)$ or $p = (x, y, z)$ or by referring to a sensor or location having a fixed position and dimension in space. A chronologically ordered sequence of positions of one moving object can be regarded as an abstract trajectory which is spatially and temporally discontinuous. For temporal aggregation of the data, time is divided into intervals. Depending on the application and analysis goals, the analyst may consider time as a line (i.e. linearly ordered set of moments) or as a cycle, e.g., daily, weekly, or yearly. Accordingly, the time intervals are defined on the line or within the chosen cycle. For spatial aggregation, it is necessary to define a finite set of places visited by the moving objects. For the aggregation, two different cases of object positions need to be distinguished:

- The object positions in the data are limited to a finite set of predefined positions, such as positions of sensors or cells of a mobile phone network.
- The object positions in the data are arbitrary. This is the case when the positions are received from mobile devices worn by the objects and capable of measuring absolute spatial positions, such as GPS devices.

In the first case, the different positions from the data can be directly used as places for the aggregation. To do analysis at a higher spatial scale, the analyst may group neighbouring positions and define places as convex hulls or spatial buffers or Voronoi polygons around the groups. In the second case, spatial tessellation may give the required set of places (space compartments) for the aggregation. Arbitrary divisions, such as regular grids, do not reflect the spatial distribution of the data. It is more appropriate to define space compartments so that they enclose existing spatial clusters of points.

However, these clusters may have very different sizes and shapes, which has two disadvantages. First, it is computationally hard to automatically divide a territory into arbitrarily shaped areas enclosing clusters. Second, the areas would differ much in their sizes, and the respective aggregates would be incomparable. Therefore, we suggest a method that divides a territory into convex polygons of approximately equal sizes on the basis of point distribution [Andrienko & Andrienko 2011]. The method finds spatial clusters of points that can be enclosed by circles with a user-chosen radius. A concentration of points having a larger size and/or complex shape will be divided into several clusters. The centroids of the clusters are then used as generating points for Voronoi tessellation [Dirichlet 1850, Voronoï 1908]. The centroids are the points with the minimal average distance to the cluster members. They are usually located inside concentrations of points.

On the basis of the defined set of places P, each trajectory is represented by a sequence of visits $v_1, v_2, \ldots, v_n$ of places from P. A visit v_i is a tuple $< o_k, p_i, t_{start}, t_{end} >$, where o_k is the moving object, $p_i \in P$ is a place, t_{start} is the starting time of the visit, and t_{end} is the ending time. Complementary to this, each trajectory is also represented by a sequence of moves $m_1, m_2, \ldots, m_{n-1}$, where a move m_i is a tuple $< o_k, p_i, p_{i+1}, t_0, t_{fin} >$ describing the transition from place p_i to place p_{i+1}. Here t_0 is the time moment when the move began (it equals tend of the visit v_i of the place p_i) and t_{fin} is the time moment when the move finished (it equals t_{start} of the visit v_{i+1} of the place p_{i+1}). It should be borne in mind that consecutively visited places p_i and p_{i+1} in a discontinuous trajectory are not necessarily neighbours in space. Having a dual representation of trajectories, as sequences of visits and as sequences of moves, the data can be aggregated in two complementary ways. First, for each place p_i and time interval Δt, the visits of this place in this interval are aggregated, i.e., the tuples $< o_k, p_i, t_{start}, t_{end} >$ where $\forall t : t_{start} \leq t \leq t_{end}$ and $t \in \Delta t$. The count of the visits and the count of different visitors (o_k) are computed. If the original data records include additional attributes, various statistics of these attributes can also be computed, such as minimum, maximum, average, median, etc.

Hence, each place is characterized by two or more time series of aggregate values: counts of visits, counts of visitors, and, possibly, additional statistics by the time intervals. The second way of aggregation is applied to $links$, i.e., pairs of places $< p_i, p_j >$

such that there is at least one move from p_i to p_j. For each link $< p_i, p_j >$ and time interval Δt, the moves from p_i to p_j in this interval are aggregated, i.e., the tuples $< o_k, p_i, p_j, t_0, t_{fin} >$ where $t_{fin} \in \Delta t$ (which means that only the moves that finished within the interval Δt are included). The count of the moves and the count of different objects that moved (o_k) are computed. If the original data includes additional attributes, it is also possible to compute changes of the attribute values from t_0 to t_{fin} and then aggregate the changes by computing various statistics. Hence, each link is characterized by two or more time series of aggregate values: counts of moves, counts of moving objects, and, possibly, additional statistics of changes by the time intervals. These two ways of aggregation support two classes of analysis tasks:

- Investigation of the presence of moving objects in different places and the temporal variation of the presence. The presence is expressed by the counts of visits and visitors in the places.
- Investigation of the flows (aggregate movements) of objects among different places and the temporal variation of the flows. The flows are represented by the counts of moves and moving objects for the links. These aggregate attributes are often referred to as flow magnitudes.

In both classes of tasks, the aggregated data can be viewed in two ways. Obviously, the data can be viewed as time series associated with the places or links. The analyst can investigate the individual time series or groups of time series (e.g., clusters of similar time series) using existing methods for time series analysis. On the other hand, the data can be viewed as a sequence of spatial situations associated with the time intervals. A spatial situation is the distribution of the object presence or flows over the whole territory during a time interval. The different views on aggregated movement data are illustrated by maps in Figure 4.5.

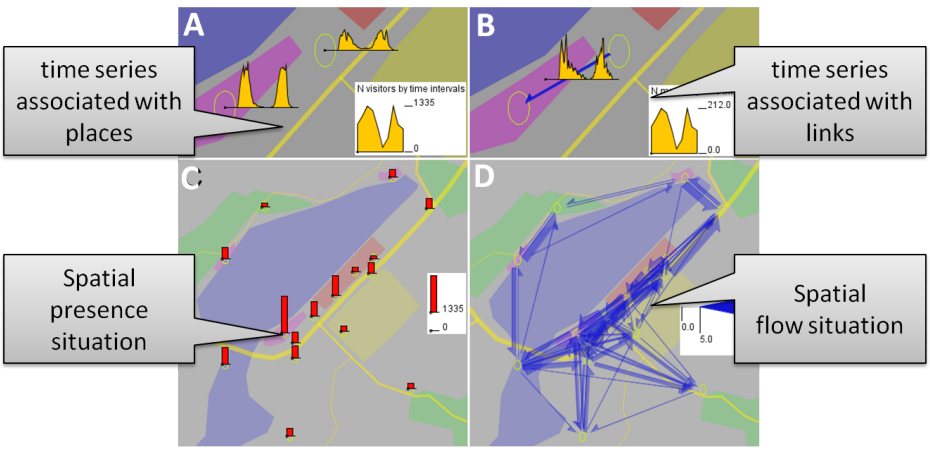

Figure 4.5: Views on Aggregated Episodic Movement Data [Andrienko *et al.* 2012].

In Figure 4.5A and 4.5B, time series of aggregate values associated with two selected places (A) and with a selected link between two places (B) are represented by polygons where the horizontal dimension represents time and the height is proportional to the values in different time intervals. The places themselves are represented

by ellipses and the link by a special symbol (further referred to as flow symbol) looking as a half of an arrow and pointing in the direction of the movement. Such half-arrow symbols are used to be able to represent flows between two places in two opposite directions. In Figure 4.5C and 4.5D, spatial situations in a selected time interval in terms of presence (C) and flows (D) are shown. The presence is shown by proportional heights of the bars drawn in the places and the flow magnitudes by proportional widths of the flow symbols. A map where aggregated movement is shown by flow symbols is called flow map [Kraak & Ormeling 2003]. It should be considered that by convention flow symbols (e.g. arrows) represent only counts of items or amounts of goods moving between some places but not the routes of the movement. In Figure 4.5D there are many intersections among the flow symbols, which clutter the display. This is a consequence of the discontinuity of the original trajectories, where consecutive recorded positions may be quite distant in space. For the brevity sake, we shall call spatial situations in terms of presence as *presence situations* and spatial situations in terms of flows as *flow situations*.

4.4.3 Clustering of Presence Situations

Clustering of spatial situations in different time intervals by similarity reduces the workload of the analyst: instead of exploring each situation separately, it is possible to investigate groups of similar situations. Besides, an appropriate visual representation of the clustering results can disclose the patterns of the temporal variation: whether similar spatial situations occur adjacently or closely in time or may be separated by large time gaps, whether the changes between successive intervals are smooth or abrupt, whether the variation is periodic, etc. This method was introduced within the joint publication [Andrienko *et al.* 2012].

For the clustering of presence situations, in each time interval the presence situations may be represented by a feature vector consisting of the presence values (i.e., the counts of visits and/or visitors) in all places. Any partition-based clustering algorithm can be applied to these feature vectors. For example we may apply the k-means clustering from the WEKA library [Hall *et al.* 2009], as in [Andrienko *et al.* 2012]. The results of the clustering are immediately visualized. The centres of the clusters are projected onto a two-dimensional colour space as shown in [Andrienko *et al.* 2012] (for convenience cited in Figure 4.6). In the upper left corner of the figure the cluster centres are represented by dots. This is done by means of Sammon's mapping [Sammon 1969]. The projection display is used for three purposes:

- first, for assigning colours to clusters so that close clusters receive similar colours,
- second, for testing the sensitivity of the clustering results to the parameters of the algorithm (k in our example), and,
- third, for detecting very close clusters that can be united.

Thus, in our example we have tried different values of k from 5 to 20 and found that starting from $k = 11$ increasing the value of k results only in appearing new dots very closely to one or more other dots while the number and relative positions of

the dots in the remaining space do not change. Closeness of dots means that the respective clusters do not substantially differ. Hence, we take the result for $k = 11$; however, it contains a concentration of five dots close to each other, i.e., the respective clusters are very similar. Decreasing k does not unite these clusters but decreases the number of the other clusters whose centres are not so close. This means that the clustering algorithm tends to produce more clusters where the data density is higher. To decrease the number of close clusters while preserving the clusters that are less similar, we apply the clustering tool only to the members of the five close clusters and set k to 2. In the result, the five chosen clusters are replaced by two clusters. In total, we have eight sufficiently dissimilar clusters of the presence situations.

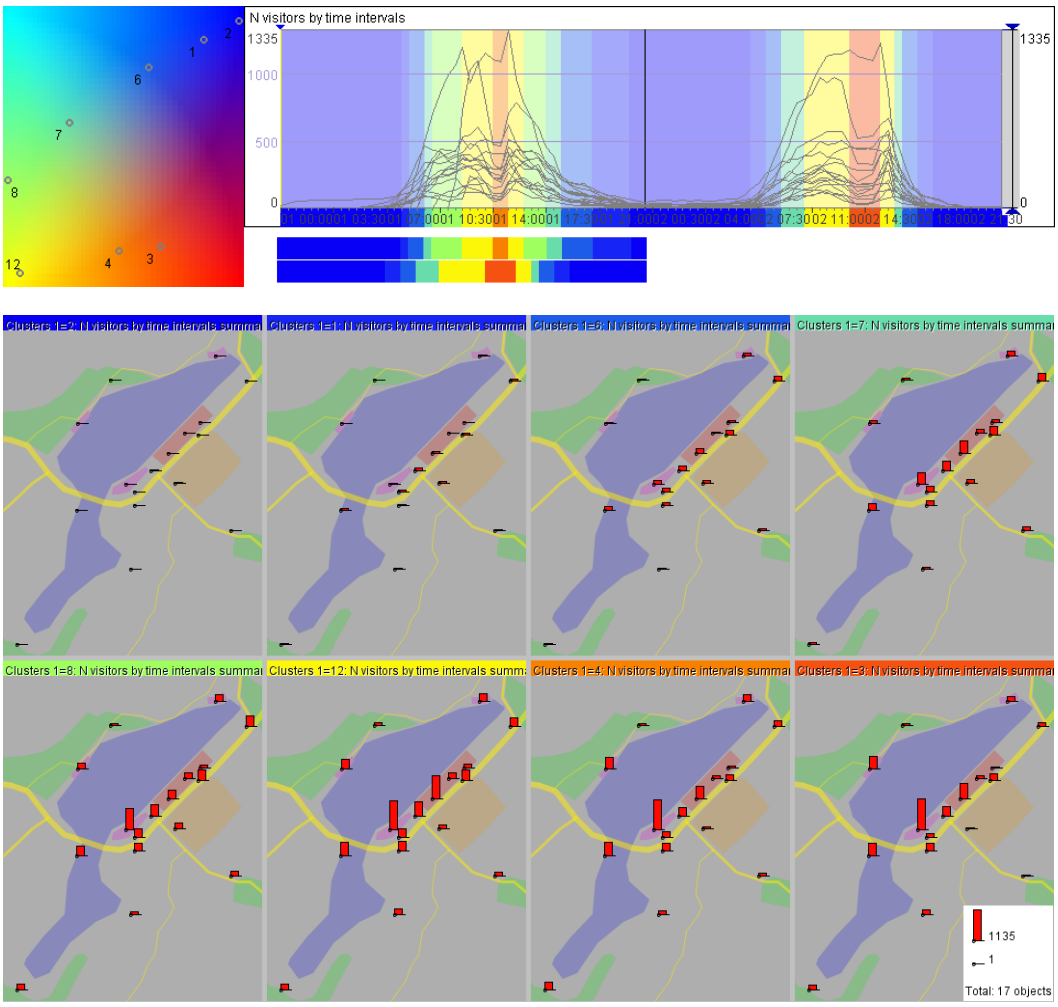

Figure 4.6: Clustering of Presence Situations. The presence situations in different time intervals have been clustered by similarity. The cluster colours are propagated to the respective time intervals. The presence situations are summarized by the clusters. The mean values of the presence are shown by proportional heights of the bars [Andrienko *et al.* 2012]

The clusters are represented in a summarized way on a multi-map display as in

Figure 4.6. Each of the small maps represents a cluster; the map caption has the colour of the cluster. To obtain a summary of a cluster, the descriptive statistics of the presence values for the places (minimum, maximum, sum, mean, median) are computed from all situations included in the cluster. One or more of these statistics can be visualized on the multiple maps. In Figure 4.6, the mean numbers of place visitors are represented on the maps by proportional heights of the bars. The colours of the clusters can be used for colouring time intervals in temporal displays, such as time graph (Figure 4.6 top right) and time mosaic (Figure 4.6 bottom), which can be used for exploring the temporal patterns of the variation of the presence situations. The time graph shows the time series of number of visitors to the places.

4.4.4 Clustering of Flow Situations

The spatio-temporal variation of the flows is explored analogously to the variation of the presence except that the clustering and visualization tools are applied to the flow situations instead of the presence situations. We proposed this analysis within the joint work [Andrienko *et al.* 2012]. The flow situation in each time interval is represented by a feature vector consisting of the flow magnitudes (counts of moves and/or moving objects) of the links in this interval. For convenience, Figure 4.7, cited from [Andrienko *et al.* 2012], exemplary shows a projection of cluster centres onto the colour space and the propagation of the cluster colours to the time graph of the counts of moves and to the time mosaic.

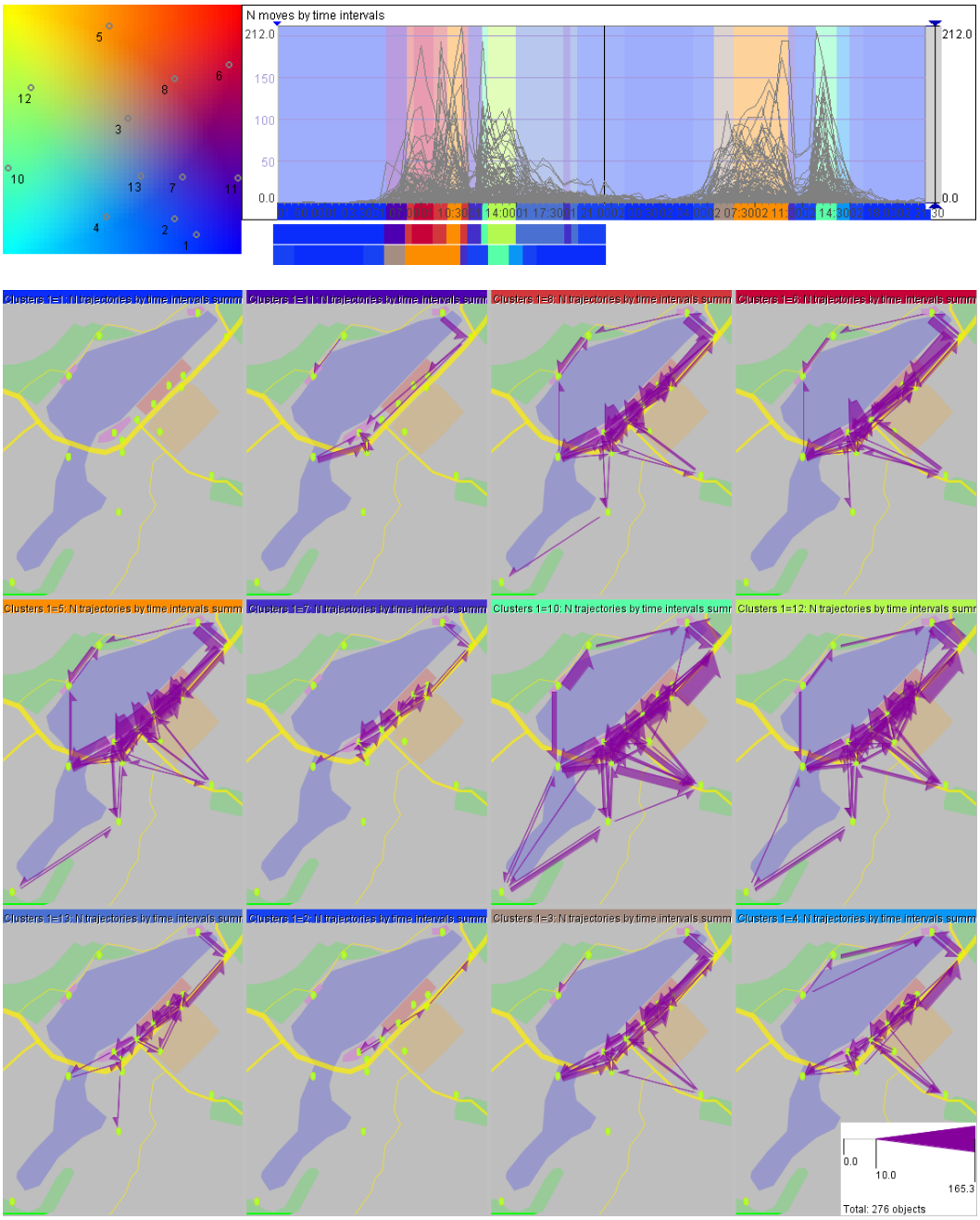

Figure 4.7: Clustering of Flow Situations. The flow situations in different time intervals have been clustered by similarity. The cluster colours are propagated to the respective time intervals. The flow situations have been summarized by the clusters. The mean flow magnitudes are shown [Andrienko *et al.* 2012]

4.4.5 Modelling Correlations with Spatial Bayesian Networks

Visual exploration of the recorded episodic trajectories gives indispensable insights on pedestrians' movement. For determination of visitor preferences or identification of potential hazards it is necessary to discover the dependencies, correlations and patterns among the movements. Therefore, this section tackles the computationally enabled visual exploration of a Bluetooth tracking dataset for inner dependencies which result from the non-random movement of the people. Existing approaches, e.g. direct database access or usage of a trajectory data warehouse (TDW) [Orlando *et al.* 2007, Raffaetà *et al.* 2011] are unfeasible, since the first one requires powerful database hosts and the second one pre-aggregates the data and so it prevents further analysis.

In this section we present a model-based approach which overcomes the limitations of existing methods by construction of an intermediate probabilistic model which preserves major location dependencies within the tracking data. Our approach represents the movement data by an easy-to-handle descriptive model, namely a *Spatial Bayesian Network* (SBN)[4]. This probabilistic model denotes the conditional probabilities among visits to discrete locations and it thus holds all required information in a compact format for further querying. Afterwards, the previously trained SBN is utilized for visual analysis and depiction of the co-visits' distribution.

Location dependencies describe the co-occurrence of geographic locations within a trajectory. They occur naturally as personal movement is purpose-driven and not a random walk through a city. Location dependencies can be expressed as conditional probability to visit an arbitrary location given that another (set of) location(s) is visited within a trajectory as well. More formally, given a finite universal set $\mathcal{L}$ of discrete geographic locations, a set $L^+ \subseteq \mathcal{L}$ containing locations that are visited with certainty and a set $L^- \subseteq \mathcal{L} \setminus L^+$ containing locations that are not visited with certainty within a trajectory, we can specify the location dependency of an arbitrary location $l \in \mathcal{L}$ by the probability $P(l \mid L^+, \neg L^-)$. The sets L^+ and L^- are also called positive and negative evidence, respectively.

The task to extract and preserve such dependencies from a dataset into a Bayesian Network is twofold (1) search for the Bayesian Network Structure and (2) assigning the common probability tables to each random variable. This task is called Bayesian Network Learning. Many algorithms tackle this task, in this work we base our analysis on the Scalable Sparse Bayesian Network Learning algorithm (SSBNL) [Liebig *et al.* 2008] as this was especially designed to meet the demands of spatial data mining. The Scalable Sparse Bayesian Network Learning (SSBNL) algorithm [Liebig *et al.* 2008] combines the advantages of the Sparse Candidate [Friedman *et al.* 1999] and the Screen Based Network Search [Goldenberg & Moore 2004]. It bounds the number of possible ancestors similar to [Friedman *et al.* 1999] by pre-sampling a given sparseness in

[4] **published with a major contribution of the author in:**
T. Liebig, Z. Xu and M. May. *Incorporating Mobility Patterns in Pedestrian Quantity Estimation and Sensor Placement*. In J. Nin and D. Villatoro, editors, Proceedings of the First International Workshop on Citizen Sensor Networks CitiSens 2012, LNAI 7685, pages 67–80. Springer, 2013
T. Liebig, C. Körner and M. May. *Scalable Sparse Bayesian Network Learning for Spatial Applications*. In ICDM Workshops, pages 420–425. IEEE Computer Society, 2008.
T. Liebig, C. Körner and M. May. *Fast Visual Trajectory Analysis Using Spatial Bayesian Networks*. In ICDM Workshops, pages 668–673. IEEE Computer Society, 2009.

the database, and bounds the edgeset to most significant dependencies by only processing frequent itemsets similar to [Goldenberg & Moore 2004]. This is done in a two-step algorithm: First, we pre-sample within each route a set of maximal k distinct locations uniformly distributed among the trajectory. Afterwards, we enumerate frequent variable sets on this pre-sampled data with threshold t and maximal length ml. The result is a bounded number of location-subsets adjustable in their size. For each of these sets a local Bayesian Network is determined in a second step that fits the original data best and the involved edges become collected on a stack. Next, this stack is sorted according to the score of the local networks. In a third step, edges are drawn from the ordered stack to construct a global Bayesian Network. Constraints for this selection are that every chosen edge must not create any cycle in the network but increase the score of the final network. Afterwards, a final database scan of the original trajectory dataset is required to recompute the common probability tables for each vertex in the global Bayesian Network.

The whole Scalable Sparse Bayesian Network Learning (SSBNL) algorithm uses pre-sampling to transform an arbitrary dataset to a processable one with adjustable size and density. Although being an approximation algorithm, the guaranteed output is one of its main advantages. It gives a reasonable approximation for positive correlations [Liebig *et al.* 2008], because the most significant dependencies persist the pre-selection of variables.

However, in order to answer queries correctly in our visual trajectory analysis, the model also needs the ability to represent negative correlations. Otherwise we are unable to express exclusive or (XOR) relations among locations in a trajectory, e.g. "If a person passes location A it is unlikely to pass location B within the same trajectory". Including edges to a Bayesian Network is always possible, if it does not create directed cycles in the network structure. Thus we sample multiple pairs of variables. In case both variables of a pair correlate negative and an edge would be valid and increases the network score, we insert an edge into the network (see lines 18 to 27 in Algorithm 2). This pairwise approach is reasonable as shown in [Meilă 1999]. The complete network learning Algorithm is summarized in Algorithm 2.

The algorithm was successfully applied to aggregated mobile phone data in [Liebig *et al.* 2009] and the application of the algorithm to Bluetooth tracking data is described in [Liebig *et al.* 2013] and presented in Section 6.4.3.3. The SBN model provides a compact and generative representation of the latent correlations within the trajectory dataset. The visual user interface, integrated into a Geographic Information System, interacts only with the model and is thus independent of the size of the underlying trajectory database.

Algorithm 2 SCALABLE SPARSE BAYESIAN NETWORK LEARNING [LIEBIG *ET AL.* 2008]

Input: D **, complete dataset**
k **, maximal frequent set size**
ml **, frequent set length**
t **, support threshold**
n **, number of random edges**
$g(\cdot)$ **, Bayesian Network score**
Output: BN **, a Bayesian Network**

1: **for all observations** $\omega \in D$ **do**
2: ω' **:= sample** k **locations from** ω
3: **add** ω' **to** D'
4: **end for**
5: FS **:= enumerate frequent sets (**D'**,**t**,**ml**)**
6: **for all** $fs \in FS$ **do**
7: $BN^* = \arg\max_{BNonfs} g(BN, D)$
8: **add edges of** BN^* **to** $edgedump$ **or if already in** $edgedump$ **increase their score**
9: **end for**
10: **sort** $edgedump$ **decreasing**
11: **for all** $edge \in edgedump$ **do**
12: **if** $BN \cup edge$ **contains no cycle then**
13: **if** $g(BN \cup edge) > g(BN)$ **then**
14: **add** $edge$ **to** BN
15: **end if**
16: **end if**
17: **end for**
18: **for** $i = 1$ **to** n **do**
19: **sample 2 different locations** X_1, X_2
20: **if** X_1, X_2 **correlate negative then**
21: **if** $BN \cup edge(X_1, X_2)$ **contains no cycle then**
22: **if** $g(BN \cup edge) > g(BN)$ **then**
23: **add** $edge$ **to** BN
24: **end if**
25: **end if**
26: **end if**
27: **end for**
28: **return** BN

4.5 Summary

In this chapter we described how to analyse Bluetooth tracking data (with visual analytics methods, by computing its representativeness, by using microscopic simulation and macroscopic analysis and by modelling inner trajectory dependencies and sequence patterns) and how to utilize episodic movement data for mobility pattern acquisition.

Hence, the representativeness of the recorded data was firstly addressed. Blue-

tooth tracking data processing records just a small percentage (about 7%) of the present people (i.e. those which carry along a mobile device with enabled Bluetooth visibility). Besides creating the awareness for this challenge when processing Bluetooth data (also stated in recent related literature e.g. [Versichele *et al.* 2012a, Versichele *et al.* 2012b, Qiang *et al.* 2012, Naini *et al.* 2011, Leitinger *et al.* 2010]), for the first time, we contribute a detailed analysis of the spatio-temporal representativeness of the recorded sub-sample. This preliminary study was conducted during a soccer match at a multipurpose arena, as the application scenarios also include a soccer stadium. The results are promising, since Bluetooth tracking provided a high correlation to the automatically recorded access control data (at lowest 0.98) for discrete spatio-temporal slices (space separated by gate, time aggregated in 15 minutes). However, this justification of utilization of Bluetooth tracking for pedestrian monitoring needs to be repeated for every application scenario, because an (eventually) varying socio-demographic composition of the persons in another closed environment, could cause different Bluetooth usage behaviour.

After this preliminary test, we recalled the distinction of microscopic mobility and macroscopic mobility from Section 2.4. We discussed methods for analysis of Bluetooth tracking data regarding both aspects, the microscopic as well as the macroscopic ones.

Firstly, for the analysis of microscopic mobility (i.e. locations and routes) from Bluetooth tracking data we described two methods. The first utilizes a state-of-the-art microsimulation (Generalized Centrifugal Force Model) and proposes the adjustment of the simulation to the recorded data. The second one applies radio signal strength fingerprinting in order to extract the individual locations and route choices. The studies [Utsch & Liebig 2012] revealed that Bluetooth tracking provides microscopic location reproduction with an accuracy of up to 4 meters and robust route choice reproduction, regarding the exits of a closed environment or the intermediate targets among the rooms.

Secondly, we focussed on the analysis of macroscopic pedestrian mobility from Bluetooth tracking data which is subsumed by the novel term *Episodic Movement Data* as it comprises the uncertainties on *continuity*, *accuracy* and *coverage*. Therefore, we considered analysis of the most frequent sequence patterns (introduced in [Kisilevich *et al.* 2010]) and our contributed methods for visual analysis of the raw data and its spatio-temporal aggregates (*move counts* and *flow counts*). Afterwards, temporal clustering of these aggregates was proposed which provide even more detailed spatio-temporal insights of the crowd movement and its phases. As the individual movement preferences and movement patterns lead to co-visit correlations among the locations we contributed a method for robust visual analysis of these dependencies which applies Spatial Bayesian Networks.

Proposed methods for Bluetooth tracking data analysis and their applications in real-world scenarios will be discussed in Chapter 6. In combination with the contributed pedestrian quantity estimation methods (previous chapter) and goal-oriented software framework (next chapter), our contribution to the Bluetooth tracking, with hardware assembling, as well as introducing specific analysis methods rounds the analysis of pedestrian mobility up.

Chapter 5

A System for Pedestrian Mobility Analysis

"It is not enough for code to work."

—Robert C. Martin [1]

Contents

Previous chapters discussed the specifics of pedestrian movement and motivated the question for pedestrian volume estimation. Empirical facts on movement behaviour have been highlighted as well as methods for digital data storage. We described how motivated individual mobility leads to movement patterns in mobility recordings and presented models for their representation. Furthermore, various pedestrian models have been discussed which describe different aspects of mobility (microscopic or macroscopic ones) based on preliminary assumptions.

[1] American software consultant and author, born 1975, Clean Code: A Handbook of Agile Software Craftsmanship [Martin 2009]

Utilizing this preliminary work, this chapter focuses on the software design of a *system* for pedestrian mobility analysis. Based on the requirements which will be elicited from the use cases, we contribute a comprehensive approach: The components: (1) user interaction, (2) empirical data recording and (3) data analysis are combined to a system for pedestrian mobility analysis.

5.1 Introduction

Pedestrian quantity estimation is a crucial task for *Location Evaluation*, *Attractor Identification* and *Abnormality Detection* application scenarios. The measurement data is limited to a bounded number of locations due to budget restrictions. Therefore, quantity estimation methods are required to estimate the number of people at unobserved locations.

A software system for pedestrian mobility analysis supports in each of the application scenarios the analyst and creates a comprehensive system to derive the needed quantity estimations. It is required to provide a system that is multi faceted and adaptable for each of the application scenarios. We contribute in this thesis not only with two efficient approaches to quantity estimation (Chapter 3) and methods for analysis of Episodic Movement Data (Chapter 4), but also with a software system for pedestrian mobility analysis in this chapter.

In this chapter we present the system and an instance of an implementation and application to real world data. Starting from the application domain perspective, we derive a model architecture and create a software solution [Bruegge & Dutoit 2010] for the pedestrian quantity estimation problem.

To come up with a software system for pedestrian mobility analysis we perform several consecutive software design steps. Next sections give preliminary definitions and a brief introduction to software engineering (for a comprehensive introduction into software development compare [Jacobson & Ng 2004]). The first step is the *requirement engineering* phase. This phase comprises (1) requirement elicitation: non-functional and functional requirements will be derived from the application scenarios. The next step is to derive use cases and user roles. After the requirements engineering phase the *architecture design* phase is tackled. The following sections discuss and present the phases of the software system development for our proposed system.

5.2 Preliminary Definitions

Before we address the software architecture process in detail, we succinctly introduce some preliminary definitions for the software design domain.

Definition 18 (Software Architecture) *The architecture of a software system consists of*

- *its structure (i.e. its components),*
- *its interfaces and*
- *relations among the components [Bass et al. 2003].*

Thus the architecture has to (1) define the components of a software system, as well as (2) its interactions and needs to (3) characterize its properties.
The architecture therefore describes static as well as dynamic aspects. It defines the blueprint as well as the operation chart. Thus, the software architecture bridges the gap between the application domain and the software implementation. The application domain is the problem from the real-world and the software implementation is the solution to it. The purpose of software architecture is described in [Starke 2005] as:

- to make software systems easier understandable
- to make a software system more flexible and reusable
- to provide an abstraction and a filtered view on the relevant things
- to ensure software quality (performance, understandability, flexibility).

Furthermore, an architecture description needs to provide answers to the questions:

- Which responsibilities does every box in the diagram have?
- Which semantic has each connecting link? At which time which information is transferred?
- For each of the connecting links: why does it exist?

Foremost, the design process of software architecture starts with the *requirement elicitation* [Starke 2005]. In this step the technical requirements of the software system are defined. This comprises

- the *functional requirements* which describe the required capabilities (e.g., the software has to provide read and write operations to the database) and
- the *non-functional requirements* which specify the required constraints (e.g., reliability, robustness, performance).

Based on these requirements and the external factors of influence the software architecture is created. Often, the requirements change during the design process [Starke 2005]. And that is why the process of the software design takes place at best in an *agile* way, by repeating the steps in a cycle [Beck *et al.* 2001] The analysis step in a software development project generates the *scope* [Starke 2005]. It ensures that the system which is developed is appropriate for its usage.

5.3 Requirements Elicitation

Described in [Starke 2005] the software system design process starts with the *requirement elicitation* step (see Section 5.2). In this step numerous features of the system are specified. Comprising (1) the *scope* of the software system, (2) how will it be used, what kind of work do the users want to do, (3) who will use it, how does it fit into the larger picture, (4) how does it affect the goals of the organization and (5) what is inside and what is outside.

We therefore expose this step for our software. A description of the possible application scenarios is provided next.

5.3.1 Application Scenarios

The application scenarios focussed in this thesis include *Billboard Location Evaluation*, *Visitor Monitoring* and *Event Monitoring*. In each of these scenarios the analysis of mobility, in detail the quantitative analysis of pedestrian movement, is highly important. Nevertheless, the scenarios differ as discussed in the following.

5.3.1.1 Billboard Location Evaluation Scenario

In the *Billboard Location Evaluation Scenario*, an analyst models the "visit potential" [Körner *et al.* 2010] of an advertisement campaign. A campaign consists of multiple billboard locations, and the "visit potential" can be computed by the pedestrian quantities [Liebig & Xu 2012].

In the indoor scenario the GPS based framework for "visit potential" estimation [Körner *et al.* 2010] is infeasible due to the lack of GPS signal. Frequency sensors may not be placed at any location, due to budget constraints and suitability of the scaffoldings. Thus, the pedestrian quantities are computed based on few sensor readings by an expert, and the resulting estimation is handed to a domain expert in order to estimate "visit potential" based on the previous model.

Thus, the pedestrian quantity estimation system has one user (the expert) who incorporates input data given by *topological data* and *sensor readings*. Furthermore, since the expert performs this task, he achieved *domain expert knowledge* which is included also in his estimation.

The expert is a domain expert, and therefore knows how to work with geographical information systems. He utilizes software tools in order to estimate the quantities at unmeasured locations given the three-tier input data (topological data, sensor readings, domain expert knowledge). The resulting model is visualised by the expert and in the eventual case of in-plausibility, the input data or the software parameters will be adjusted. The resulting model can be stored in a generic transferable format, which is not just understandable by this expert user but by any unexperienced user as well.

However, the user role that interacts with the software system in the billboard location evaluation scenario is just *the domain expert*. As derived from the description above, the following functional requirements are posed to the system in this scenario:

- The system must compute quantity estimates.
- The system should use topological floor plan information.
- The system should not depend on high granular maps.
- The system must incorporate sensor readings.
- The system should support seamless integration of various sensor technologies (Bluetooth and other methods from Section 2.3).

as well as the non-functional requirements:

- The system must provide results in a re-usable format.
- The system has to be robust.

5.3.1.2 Visitor Monitoring Scenario

In the *Visitor Monitoring Scenario*, the visitors of a zoo or a fun park are subject for analysis. For the zoo, the knowledge on the attractiveness of the compounds and shops is very interesting. This can be measured either in the quantity of persons or their stay times at various locations in the zoo. In order to evaluate and plan the signage of the zoo, it is also important to analyse the sequence patterns and find a typical walk-through. However, a system for pedestrian quantity estimation is just a small part of the whole mobility analysis required for visitor monitoring in a zoo. The analysis needs to be possible without expert knowledge on geographic information systems.

User roles interacting with the software system in the visitor monitoring scenario are *the analysts*. The following functional requirements are posed to the system in this scenario:

- The system must compute quantity estimates.
- The system should use topological floor plan information.
- The system must incorporate sensor readings.

as well as the non-functional requirements:

- The system must provide results in a re-usable format.
- The system must have independent time complexity from number of monitored pedestrian movement.
- The system has to be robust.

5.3.1.3 Event Monitoring Scenario

The *Event Monitoring Scenario* is integrated in the Emergency Support System [ESS 2010]. This Emergency Support System is a software tool that supports forces in crisis management. Whenever an incident happens, the system is fast deployable and provides a comprehensive information system inside the command post. In order to deal with different incidents the system may use various technologies for data transmission including Wireless Local Area Network (WLAN), General Packet Radio Service (GPRS) and satellite communication. The different sensor readings from these heterogeneous networks are integrated in the system. The domain expert (a so-called technical commander) operates the system and depicts the sensor readings on a map. Additionally, the locations of the forces are also depicted on the map. Several expert tools are included in the web-based ESS portal via RESTful services (discussed in Section 5.5.3 and introduced in [Richardson & Ruby 2007]). These provide data analysis and prediction capabilities to the expert. One of these services is the hereby presented system for pedestrian presence analysis (called *Historical Data Analysis module* within the ESS project [ESS 2010]).

Therefore, the user role involved in this event monitoring scenario is the analyst, only. Since he does not directly interact with the software system but uses an intermediate user interface (the emergency support system) the requirements in this case are as follows. Functional requirements:

- The system must compute quantity estimates.
- The system should support seamless integration of various sensor technologies (Bluetooth and other methods from Section 2.3).
- The system must act as a RESTful service.

Non-functional requirements:

- Possible integration in real-time systems.
- The system must be integrated in a web platform (as RESTful service).
- The system has to be robust.

5.3.2 Functional Requirements

The application scenarios of the software system were discussed in detail in previous sections in order to elicit the functional and non-functional requirements posed by the applications to the software system. This requirement elicitation is the first step in software design. In summary of the industrial application scenarios, the selected requirements for the comprising quantity estimation system are:

- (F1) usage of topological floor plan information instead of high-granular maps,
- (F2) incorporation of sensor measurements,
- (F3) incorporation of expert knowledge given as movement patterns,
- (F4) estimation of traffic quantities for unmeasured locations
- (F5) independent time complexity from number of modelled pedestrians.

Next sections discuss how the hereby introduced system for pedestrian monitoring addresses the posed requirements. As previously described in Section 5.2, additionally to the functional requirements to the system the application scenarios pose non-functional requirements, which are discussed next.

5.3.3 Non-Functional Requirements

Besides the functional requirements to the software system the presented application scenarios pose non-functional requirements, which need to be taken under consideration during the software design process.

- (NF1) Seamless usage of heterogeneous sensor technologies,
- (NF2) Possible integration in real-time systems (implementation requirement).
- (NF3) Integration of monitoring results in subsequent analysis and reporting tools (implementation requirement).
- (NF4) The system has to be robust.

The chapter proceeds with formalizations of the software system including its architecture and interface description depicted in sequence diagrams. Industrial showcases of the presented system are described within the next chapter.

5.4 Architecture

The systems architecture consists of three main components:

- the *Query Layer*,
- the *Sensor Layer* and
- the *Analysis Layer*

which provide the necessary functionalities (compare sections above) to the data analyst. The overall structure is depicted in Figure 5.1. The *Query Layer* divides into the *graphical user interface* and the *controller* which triggers the processing workflow[2].

Next, we are introducing each of the three components separately and afterwards describe their interfaces and relations. According to the definition of the *software architecture* (given in Definition 18), this component diagram (Figure 5.1) describes the software architecture [Bass *et al.* 2003].

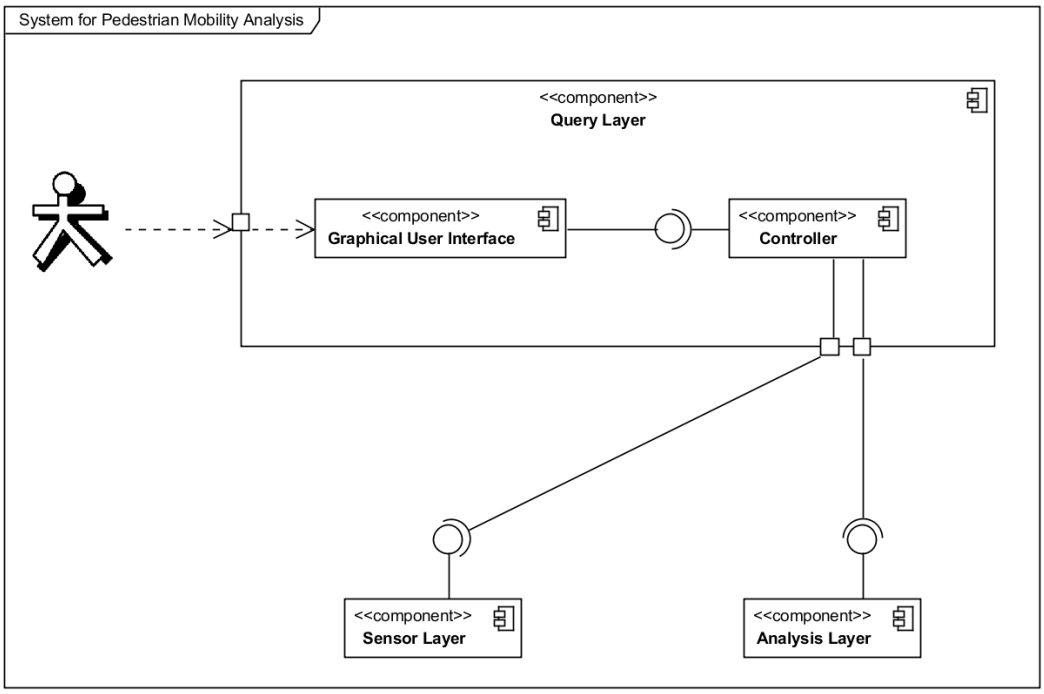

Figure 5.1: Component Diagram for the Pedestrian Mobility Analysis System.

5.4.1 Sensor Layer

The task of fetching empirical data on people's presence or movements is conducted in the *Sensor Layer*. The purpose of the sensor layer is to provide the data of multiple arbitrary sensors through a unique, standardized and open interface.

[2] **published with a major contribution of the author in:**
T. Liebig and Z. Xu. *Pedestrian monitoring system for indoor billboard evaluation*. Journal of Applied Operational Research, vol. 4, pages 28–36, 2012.

Possible pedestrian quantity sensor technologies are highlighted in Chapter 2. These comprise manual measurements, video surveillance, 3D laser scans and Bluetooth tracking. However the data is recorded, the retrieved values are transferred Open Geographic Consortium (OGC) compliant and stored in a database for later consideration, when the *Query Layer* (respectively the user) asks for creation of pedestrian models based on historical data.

5.4.2 Query Layer

In the pedestrian monitoring system the second layer of the architecture is represented by the *Query Layer*. This has the purpose to pass user-triggered analysis parameters to the *Analysis Layer* (described in nest section). The Query Layer acts twofold. It firstly provides an interface to external applications or to the analysis expert through the graphical user interface. The expert or respectively the external application might decide for a specific spatio-temporal interval for the data analysis to be performed. Secondly, the *Query Layer* acts as a controller for the whole architecture. The specified spatio-temporal interval triggers the *Analysis Layer* to perform the data analysis, incorporating latest sensor readings of the *Sensor Layer*, and defines its most important parameters. These can be space and time parameters as well as the choice of analysis algorithm.

The results may be directly be integrated in external applications and also be depicted for the analyst on a map within the graphical user interface.

The *Query Layer* supports real-time pedestrian monitoring solutions as well. The *Query Layer* regularly triggers the data analysis with the latest incoming sensed data.

5.4.3 Analysis Layer

The computation performed by the *Analysis Layer* is triggered by the previously described *Query Layer*. Thus, it includes latest sensor readings of the *Sensor Layer* and performs quantity estimation for unobserved segments of the traffic network. The considered spatio-temporal boundaries of this estimation as well as the estimation methods are previously specified by the *Query Layer*.

Few of the pedestrian models presented in Chapter 3 require detailed representations of the accessible space, but as we are only interested in quantities per location, we do not require such a detailed model and these methods are not supported by the *Analysis Layer*.

Therefore, a directed graph approximation of the floors, stairways and junctions contains enough information for our task. Every junction is represented by a vertex and the connecting floors are represented by edges.

The presented models focus on aggregations of *pedestrian presence* or of *pedestrian moves* and thus may disregard temporal aspects of movement. Compare previous chapters for more details on pedestrian quantities, traffic network construction (Chapter 2) and quantity estimation methods (Chapter 3).

5.4.4 Interface Description

Whereas previous sections describe the components of the architecture individually, this section describes their links in more detail.

5.4.4.1 Query Layer - User Interaction

The *Query Layer* poses a graphical user interface to the analyst. As there may be various roles of analysts the communication includes an authentication which may restrict the applications capabilities depending on the user's identifier. In order to keep the architecture flexible and easily applicable the graphical user interface is provided as a web-interface which can be easily accessed by a browser.
In this graphical user interface the user specifies:

- which sensor data should be incorporated,
- the placement of the sensors,
- which movement patterns should be incorporated,
- which traffic network reflects the topological floor plan information,
- which spatial boundary is the area of interest for analysis,
- which temporal interval is considered for traffic frequency aggregation and
- which algorithm should be preferably used for analysis.

As mentioned above, questions of this user dialog can be omitted depending on user role or on application context.

After specifying these parameters, the user triggers the computation. In turn, the results are returned to the user's browser using OGC compliant map interface protocols (Web Feature Service WFS).

For further use, the results of the analysis are also handed back to the user using OGC compliant storage protocols (either Geographical Markup Language GML for indoor usage or Keyhole markup Language KML for WGS 84-referenced outdoor data [National Imagery and Mapping Agency 2000]).

5.4.4.2 Sensor Layer - Query Layer

The *Sensor Layer* hands out aggregations on people's presence or moves to the *Query Layer*. This retrieval of sensor data could be performed either online or offline. Offline data collection implies batch processing of the collected data, whereas online data transmission enables also real time data processing. However, as the analysis described in this framework is triggered by the user, respectively the *Query Layer* (described in next section) both scenario (online and offline processing) are possible, depending on the computation frequency.

Nowadays, most Geographic Information Systems provide interfaces for the Keyhole Markup Language (KML) which is standardized by the Open Geographical Consortium (OGC) and may hold the spatio-temporally coded sensor readings for offline batch processing.

For online data transmission a common open protocol is the so-called sensor observation service (SOS)[3], which is a XML standard for data collection from heterogeneous mobile sensors. Core functions of this Open Geographic Consortium (OGC)[4] standard are:

- **GetCapabilities -** used in order to achieve information on the monitored value,
- **GetObservation -** used to fetch measurements and
- **DescribeSensor -** returns ID and unit of observed value.

5.4.4.3 Analysis Layer - Query Layer

After the *Query Layer* collected every required information in order to trigger computation, it hands the following parameters to the *Analysis Layer*:

- sensor readings per considered time interval,
- situation of the sensors,
- movement patterns,
- traffic network,
- spatial boundary of the area of interest for analysis,
- considered analysis method.

The *Analysis Layer* performs the computation while the *Query Layer* waits for its response. In return the *Analysis Layer* sends a list of street segments (subset of the traffic network) with the estimated values back to the *Query Layer*.

5.4.5 Sequence Diagram

Next we describe previously mentioned interaction of the components in detail for the *quantity estimation* use case, which is common in all previously introduced application scenarios. The interaction of the different components for this particular use case is depicted in temporal sequence (top to bottom) in Figure 5.2.

- As revealed in previous sections, the user starts his analysis with a registration to the system. The system (respectively the *Query Layer*) may depend in its behaviour according to the user identifier in order to provide different user roles.
- After the user logged himself to the system, he triggers the computation through the graphical user interface. The *Controller* of the *Query Layer* fetches the sensor readings for considered spatio-temporal interval and forwards the request for computation with all required parameters to the *Analysis Layer*.
- The *Analysis Layer* computes its output based on the retrieved input data and hands analysis results back to the *Query Layer*.

[3]http://www.opengeospatial.org/standards/sos, last accessed 30 September 2011
[4]http://www.opengeospatial.org, last accessed 30 September 2011

- The Controller contained in the *Query Layer* performs post-processing and hands results back to the user.
- The analysis can be refined or repeated until the user signs out from the system (i.e. Logout message is send).

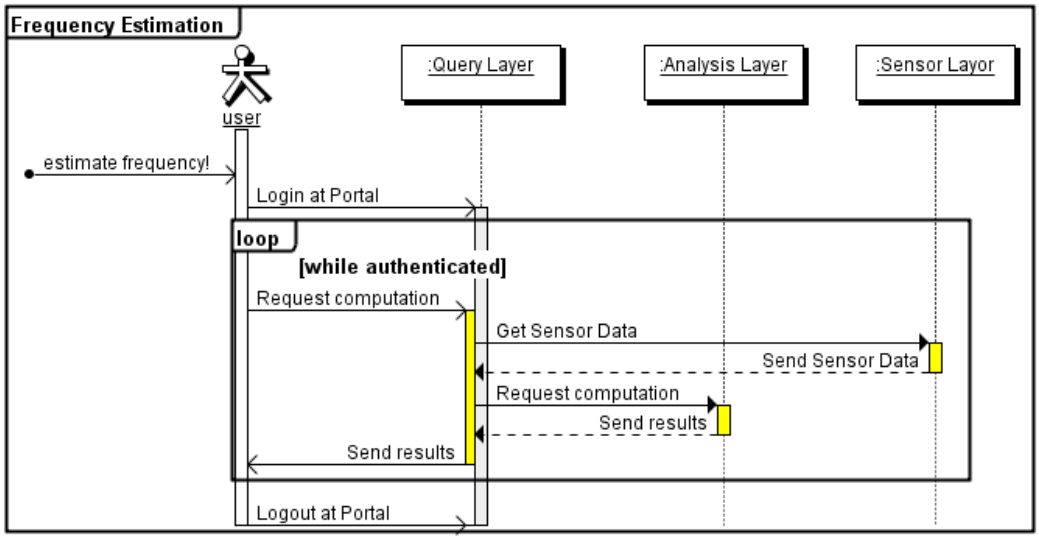

Figure 5.2: Sequence Diagram for Frequency Estimation.

5.5 Software Integration

Since previous sections focussed on the *application domain* and the software architecture, the section at-hand focuses on the *solution domain* (see [Bruegge & Dutoit 2010] for *object-oriented modelling*) and the application context. Namely, we are going to present a software integration of previously presented architecture which meets the posed requirements.

Additionally, the system performs map matching of the sensor data to a predefined traffic network, which is used for quantity estimation later on.

The steps performed after the user triggers the computation via transmission of a KML file that denotes spatial position of the sensors as well as their readings (as comma separated values) are:

- **Map Matching -** The sensor positions are matched to a given NAVTEQ®[5] traffic network using the closest segment for each segment.
- **Stencil -** Based on the bounding box of the resulting sensor positions, a subgraph for frequency estimation is obtained from the original NAVTEQ® network by incorporating all segments within a buffer of the bounding box.

[5]Usage of any other traffic network is seamlessly supported if provided as relational table of polylines, compare Section on spatial data base management systems, Section 2.2.3.

- **Transformation -** Since the quantity estimation methods (described in Chapter 3) make use of line graphs instead of NAVTEQ® street segments, the data is transformed such that it fits to the previously described methods.
- **Quantity Estimation -** The quantity for unobserved segments is calculated by one of the methods described in Chapter 3.
- **Back-Transformation -** The calculated frequency estimations are back transformed.
- **Visualisation -** The frequency estimates are depicted on a map to the user. The street segments are coloured according to the estimated traffic quantities.
- **Reuse -** The same value are provided for download as Keyhole Markup Language KML file.

The system could easily be coupled with heterogeneous sensor readings, see Chapter 7 for a discussion on future work.

5.5.1 Robustness Analysis

Main actor of the frequency estimation use case is the analyst (i.e. modelling expert). The input data required to estimate the traffic quantities (and therefore to build the mobility model) include:

- Information about the presence or moves of people. The data must have the format: $< p, y_p, t >$, where p denotes the location, y_p the aggregate *number of moves* or *number of flows* (compare Chapters 2 and 3) for the considered location,
- Time interval for analysis,
- The sensor positions in WGS 84 [National Imagery and Mapping Agency 2000],
- The traffic network in WGS 84 consisting of street segments[6],
- Movement patterns as sequences of street segment identifier and
- Additional parameters for the estimation method.

The output of the estimation is

- A KML file holding estimates and responding street segments,
- the same as WFS layer provided for map display.

A robustness analysis checks whether the usage of the system leads to unexpected behaviour. The robustness diagram unsheathes the possible courses of usage as well as occurrences of error messages [Starke 2005]. The robustness diagram for the quantity estimation is depicted in Figure 5.3. Whatever error occurs (depicted in red), either in computation or by user interaction is fetched and displayed to the user.

[6] In our implementation NAVTEQ® is the default traffic network, however it can be seamlessly replaced by any other map service stored in a relational database management system.

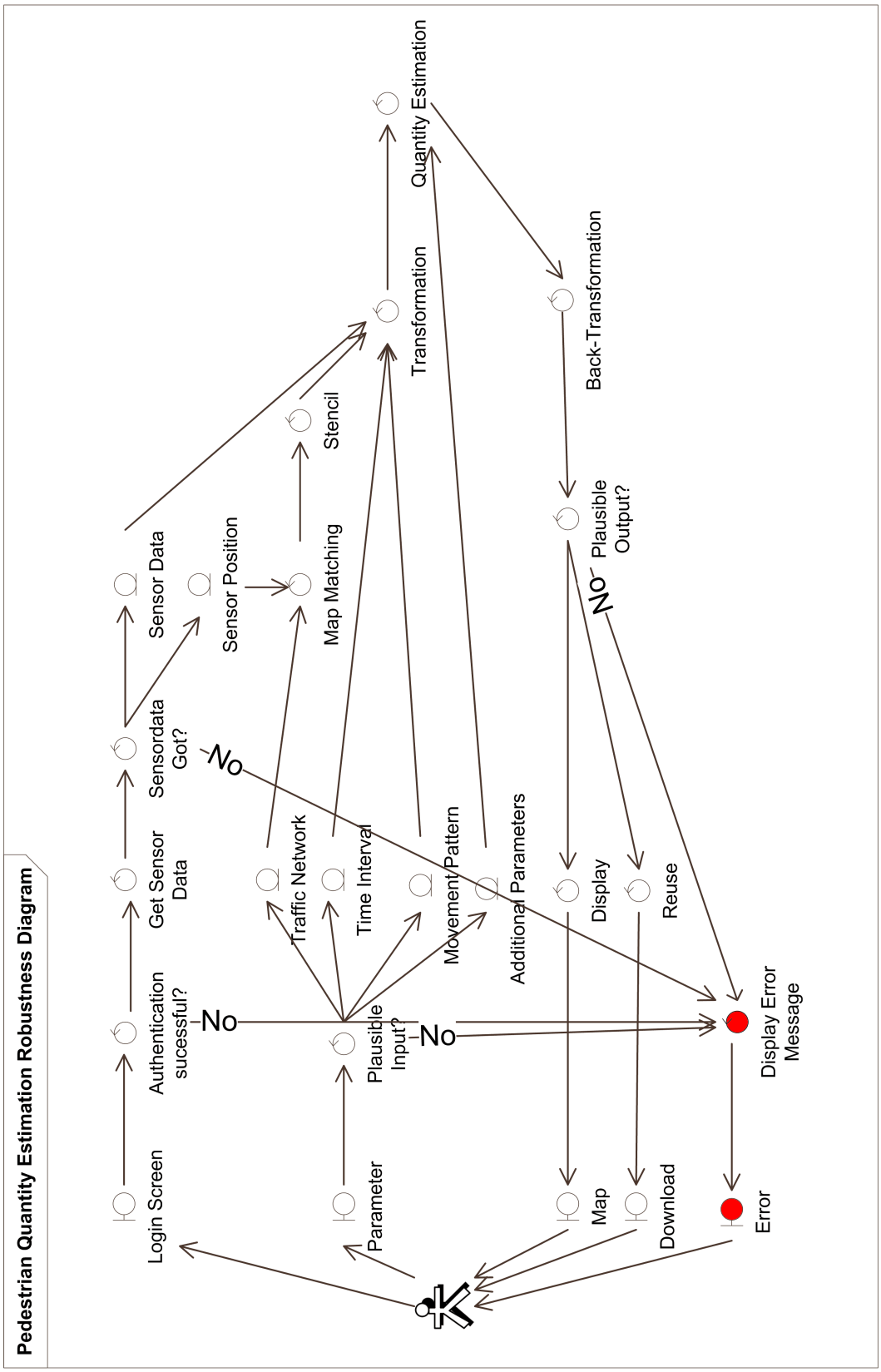

Figure 5.3: Robustness Diagram for Frequency Estimation.

5.5.2 Graphical User Interface

As described in previous sections the software implementation provides a browser interface. This allows simultaneous usage for multiple users and different roles, furthermore easy accessibility. Upcoming Figure 5.4 depicts two screens of the user interface. In the top the setting of parameters is depicted whereas the image in the bottom depicts the resulting map, including the button for KML download.

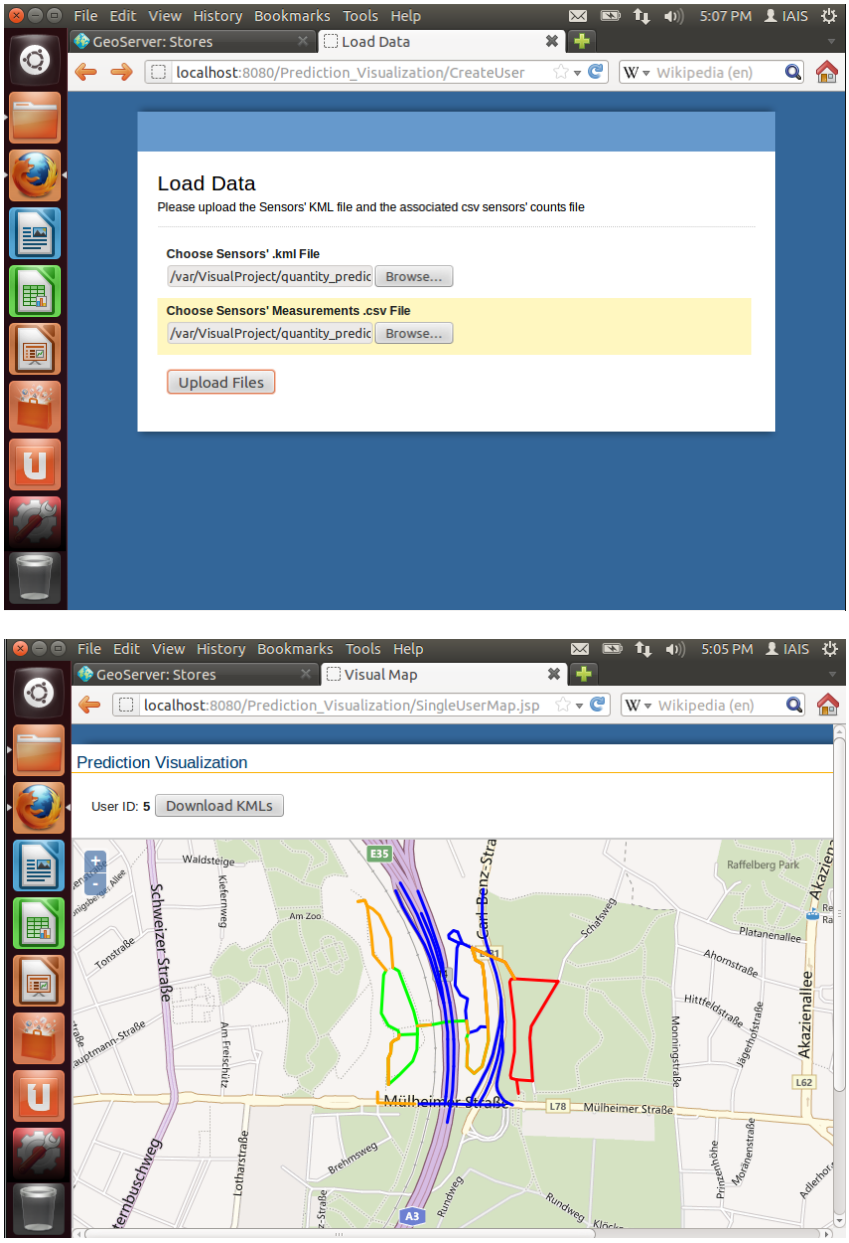

Figure 5.4: Web-based Graphical User Interface of the Software System.

5.5.3 Integration in the Emergency Support System

The hereby presented system was integrated in the web-portal of the European Support System [ESS 2010]. Scope of the project is the support of an incident commander in a command post after an incident (for example a flood or a truck crash) occurred.

Besides depiction of current and historical sensor readings and geo-positions of the forces, victims and vehicles, the system provides possibilities to send commands to the forces and to broadcast information to the people within the hazardous zone (using IMSI catcher) or to public using social media.

Furthermore, external prediction modules are connected to the system as RESTful web-services (see [Richardson & Ruby 2007] for a comprehensive introduction to this architecture). This stands for the REpresentation State Transfer architecture which contains constraints for the developer. Briefly, these are listed next, for a comprehensive introduction we refer to [Richardson & Ruby 2007].

- **client-server** An interface separates clients from service provider.
- **stateless** The server does not store a state for the communication with the client
- **accessibility** Every provided functionality needs to be directly accessible by the client.
- **predefined operations** The communication uses predefined operations for communication: `GET`, `POST`, `PUT`, `DELETE`.
- **uniform interface** The output of the server can have different formats according to the clients requirement, e.g. WFS or WMS map tiles for depiction in a map client.

One of these so built modules is the hereby described software system called 'Historical Data Analysis (HDA) Module' within the Emergency Support System.

For this integration, the graphical user interface of our system is replaced by a HTTP interface and an XML based communication protocol. The ESS Portal provides an integrated user interface with the necessary controllers and displays (map and charts) to the analyst. HDA provides various predefined models to the user depending on their meta-data. The user-chosen parameters: model selection, temporal interval and spatial boundaries are forwarded to our system, which behaves as described above and returns the estimations back to the portal in KML format. Estimations could be performed as real-time simulations or data aggregation depending on the selected model parameter. Furthermore, the HDA provides access to non-pedestrian, but vehicular data analysis using the same interface.

The main goal of this application is the estimation of the presence of people or moving vehicles in a given set of places at a given time moment or over a time interval. This may be needed for planning the distribution of the resources, evacuation, and, possibly, other emergency management activities. The foreseen functions of the application include the following capabilities:

- Estimate the presence of people in different places at a particular moment of time.

- Estimate the movement flows of people or vehicles over a territory at a particular moment of time.

The application builds and uses a prediction model based on historical data. Two kinds of models are foreseen, depending on the spatial scale and character of the area for which the modelling is done: (a) large-scale modelling for a geographical area like a district of a city, a city, or a region; (b) small-scale modelling for a building, a campus, a train station, or an airport. The work of the application is divided in two phases. In the first phase, an expert interacts with the application to build a prediction model. In the second phase, the model that has been built is used. There are two possible uses of the model. One possible use is to obtain estimated values of presence or movement for a given area and time moment or interval. The other possible use is to compare the values predicted by the model with the real values. In case of a large deviation of a real value from a predicted one, the user must be alerted. This may mean that either an exceptional situation has occurred and therefore the model's prediction is no longer valid and should not be used or that there is an error in the input data. In the phase of model building, the user (modelling expert) needs to interact with the application for choosing the right parameters, checking intermediate results, and refining the model in an iterative way until sufficient prediction accuracy is gained. In the phase of model use, the end user, such as a technical officer, does not need to interact with the application. The user sends a request with the necessary input (e.g. place and time) and receives the result (e.g. the predicted values). Since the historical data gradually grows with the real time data, the prediction model will be updated from time to time. This means that the application either rebuilds the whole model or it is incrementally updated with the incoming real time data. The modelling is based on historical data about the spatial and temporal variation of the presence of people over the area. Examples of such data are GNSS (Galileo/GPS) or RFID tracks of people, records from movement sensors in buildings, and records from traffic sensors (in case when presence or movement on roads is modelled). There may be cases when available historical data do not represent the whole population of the area. For example, data about the usage of mobile phones may be useful for the estimation of the population dynamics patterns. However, this data alone does not give the right estimation of the number of people in a given place at a given time moment because not all people use mobile phone services of a particular company. Therefore, for building a model of people presence, this data needs to be combined with other kinds of data such as population census data.

The Historical Data Analysis Module provides two use cases related to the thesis at-hand:

- Model Building
- Getting Prediction

Next sections describe these use cases in more detail.

5.5.3.1 Use Case 1: Model Building

This use case describes the steps for building a model that will be used for the estimation of people's presence or flows. Typically it is conducted by the *modelling expert*.

Main actor: modelling expert

First the data prediction model is generated using historical data as shown in Figure 5.5. The user needs to authenticate at the ESS portal. There the model building process is triggered by the user and the historical data analysis module becomes active. Afterwards, data is fetched by the HDA and used for model creation. The resulting model is stored in the data analysis module, whereas its meta-description is forwarded to the user. Utilising the model is described in the next use case.

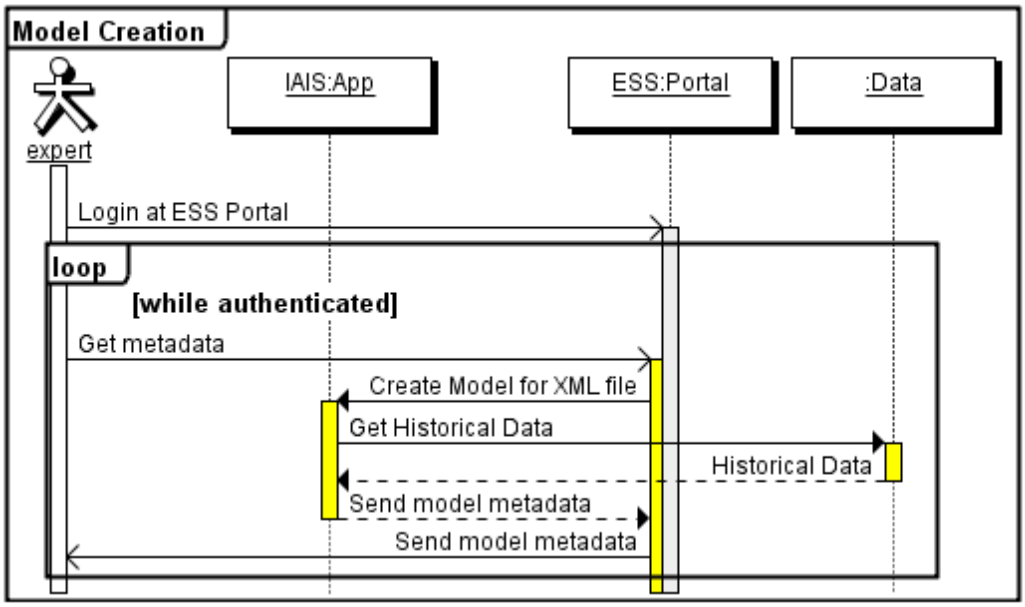

Figure 5.5: Model Building Sequence Diagram [ESS 2010].

The input data required to build the model include:

- Historical Data: Information about the presence of people or cars. The data must have the following format: $< x, y, t, v >$, where x, y are the geographical coordinates, t the time stamp of the measurement, and the value v is a quantitative measure like count of people or cars. The data should be retrievable by the application via a connection to internal and external data sources. The application shall send a request for a selection of the historical data. The request must be an ordinary SQL query on the historical data that retrieve a table of the format (x, y, t, v).

- Additional data (when necessary): population census data or other data that may complement historical data when they are incomplete.
- Selected area: the area for which the model needs to be built divided into spatial compartments (regular or irregular grid) according to the desired spatial resolution. The division may be produced by the expert in interaction with the application. It is also possible to use an existing XML/KML file defining the boundaries of the area and the compartments.
- Temporal resolution: the desired temporal resolution of the model, e.g. 10 minutes, 1 hour, 1 day, etc. This is a pair (time unit, value), which is specified by the expert in interaction with the application.
- Additional parameters: specific parameters required by the model builder. The parameters are specified by the expert in interaction with the application.

The output of the model building phase:

- An internal representation of the model stored in the internal database of the application.
- Metadata about the model. The metadata are sent to the ESS portal and stored there. This will allow end users to use the model (see use cases 2 and 3). The metadata include a unique identifier of the model, specification of the type of data predicted (number of people, number of cars, etc.), specification of the area, temporal resolution and possibly other fields.
- XML/KML file with the area and its division into compartments. Each compartment has its unique identifier. This file may later be used for allowing the end users to choose a part of the territory (one compartment or a subset of compartments) for which the prediction is required.

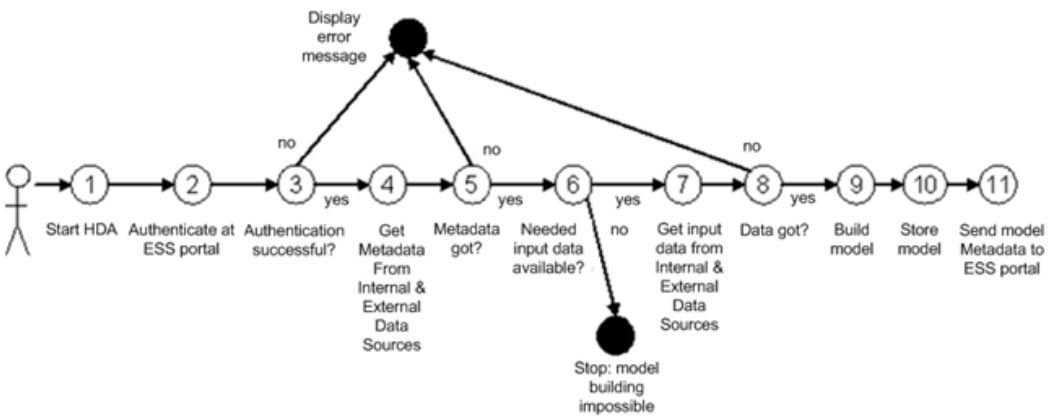

Figure 5.6: Model Building Robustness Diagram [ESS 2010].

Basic course

- The modelling expert starts HDA. HDA tries to authenticate at the ESS portal. For this purpose, it requires the user to provide the login information and sends this information to the portal.
- HDA receives a response from the ESS portal and checks if the authentication was successful. HDA sends a request for the metadata about the available historical data from the internal and external data sources (the user needs to know what historical data are available).
- HDA receives the metadata about the historical data from the internal and external data sources.
- The user inspects the metadata and finds information about the historical data required for building the model.
- HDA sends a request for the historical data required by the user to the internal and external data sources.
- HDA receives the requested historical data from the internal and external data sources.
- The user builds a model in interaction with HDA. If necessary, the user loads additional data, e.g. data available locally.
- The model which has been built is stored in the local database of HDA for further use.
- HDA sends the metadata about the model to the ESS portal.

Alternate courses

- The authentication failed. HDA displays an error message.
- The metadata about the available historical data could not be got from the internal and external data sources. HDA displays an error message.
- By inspecting the metadata, the user finds out that data suitable for building the model are not available at the internal and external data sources. The user stops the work since building a model is impossible.
- The historical data required by the user could not be got from the internal and external data sources. HDA displays an error message. This is notified to the ESS portal.

5.5.3.2 Use Case 2: Getting Prediction

This use case describes an operator using a previously built model for the estimation of people presence or other values relevant to the emergency situation.

Main actor: operator (technical officer)

After the model is created, predictions are requested by the user as shown in the following sequence diagram (compare also Figure 65 in D6.1). After authentication at the ESS portal the user requests a list of models and their metadata. The requests are forwarded to the HDA, which generates the replies. Returning the list of available models in a first step, the user continues sending detailed prediction requests through the portal to the IAIS HDA application. The requests are again forwarded to the HDA and are answered using the pre-computed models. Replies are shown to the user by the ESS portal.

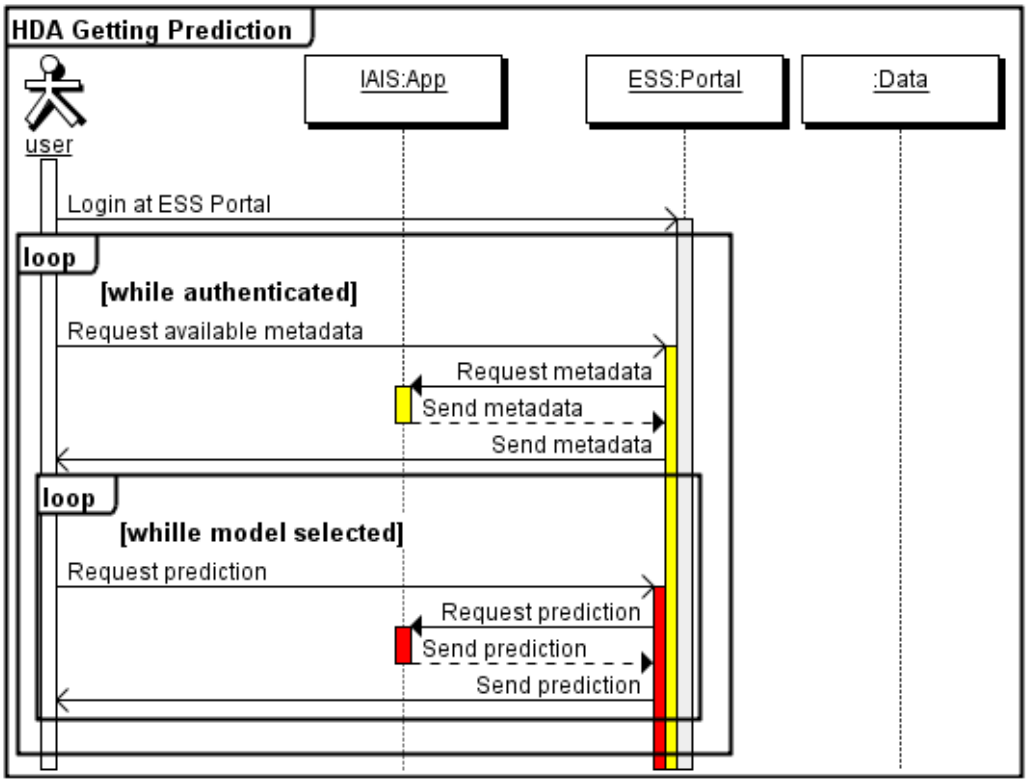

Figure 5.7: Getting Prediction Sequence Diagram [ESS 2010].

The input data required for getting a prediction is:

- The unique identifier of the model should be used. This identifier is taken from the metadata generated in the phase of model building. It is assumed that the

user views the metadata records about the existing models and selects a suitable model according to the current needs.

- Selected set of places: if the user does not need predicted values for the whole area but only for a subset of spatial compartments, the user specifies the compartments of interest. This is a list of identifiers of the compartments.
- Selected time interval: the time interval for which the prediction needs to be made. This is a pair (t_1, t_2) where t_1 and t_2 specify the start and end dates and times of the day.

The output of the application is:

- Predicted values: estimated values of presence or flows in the selected spatial compartments by time intervals according to the temporal resolution of the model. This is a KML file with the coordinates of the spatial compartments and the respective time series of the predicted values. This enables the visualisation of the results in a browser or a web map service. The data is sent to the ESS portal for communication to the user.

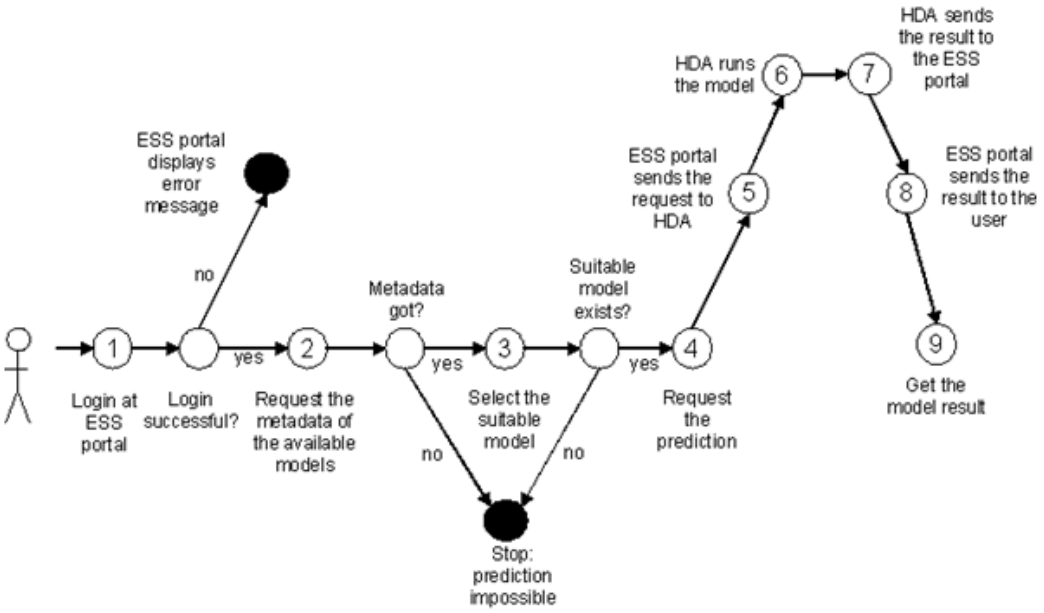

Figure 5.8: Robustness diagram for *Getting Prediction* [ESS 2010].

Basic course

- The user (operator) logs in at the ESS portal.
- The user requests the metadata about the available models (previously prepared in Use Case 1).

- The user selects the suitable model from the available models according to the type of predicted values, territory covered, division of the territory, and temporal resolution. This information is provided in the metadata.
- The user sends a request to the ESS portal for making the prediction. In the request, the user specifies the identifier of the model, the set of territory compartments (optional), and the time moment or interval for which the prediction needs to be done.
- The ESS portal sends the request to HDA.
- HDA runs the model and obtains the predicted values for the specified territory compartments (or for all compartments if the user did not select any compartments) and for the specified time unit or sequence of time units.
- HDA produces a KML file with the result of the prediction and sends it to the ESS portal.
- The ESS portal sends the result of the prediction to the user.
- The user gets the result and uses it in the following tasks (e.g. evacuation planning).

Alternate courses

- The login failed. The ESS portal displays an error message.
- There is no metadata, which means that no models have been generated so far. The user stops the work since getting a prediction is impossible.
- Among the available models described by the metadata there is no model suitable for the user. The user stops the work since getting a required or valid prediction is impossible.

5.5.3.3 Interfaces Protocols

The protocols for interfacing the RESTful service are XML based. Successful queries are answered by a message containing geometries and associated values. This communication applies the OGC compliant KML protocol. For details on the protocols we refer to Appendix A.

5.6 Summary

The chapter at-hand addressed the applicability of the previously presented methods for quantity estimation (see Chapter 3). Therefore, a system for pedestrian mobility analysis was developed. This included the requirement elicitation step, the architecture specification and the description of the implementation.

The requirements have been derived from the three application scenarios introduced in Chapter 1: Billboard Location Evaluation, Visitor Monitoring and Event

Monitoring. We derived functional and non-functional requirements based on the scenario description. The functional requirements comprise:

- (F1) usage of topological floor plan information instead of high-granular maps,
- (F2) incorporation of sensor measurements,
- (F3) incorporation of expert knowledge given as movement patterns,
- (F4) estimation of traffic quantities for unmeasured locations,
- (F5) independent time complexity from number of modelled pedestrian,

whereas the non-functional requirements are:

- (NF1) Seamless usage of heterogeneous sensor technologies,
- (NF2) Possible integration in real-time systems,
- (NF3) Integration of monitoring results in subsequent analysis and reporting tools.
- (NF4) The system has to be robust.

We met these requirements in the solution domain by a three-tier structure consisting of *Query Layer*, *Analysis Layer* and *Sensor Layer*. The *Sensor Layer* (1) monitors geo-coded sensor recordings on people's presence and hands in this episodical movement data as input to the next layer. By use of standardized OGC interfaces for data collection, we seamlessly integrate various sensor technologies depending on the application requirements. The *Query Layer* (2) interacts with the user, who could ask for analyses within a given region and a certain time interval. Results are returned to the user in OGC conform Geography Markup Language (GML) format. The user query triggers the (3) *Analysis Layer* which utilizes the mobility model for traffic volume estimation.

In the solution domain the software realization was performed by an implementation as web service. The graphical user interface is provided through the users browser. We depicted some screens for the visitor monitoring application. Furthermore, robustness of the system was analysed in a robustness diagram for the major use case.

Next chapter gives more details on analysis performed in the industrial application scenarios and presents the results in detail.

Chapter 6

Real World Application Scenarios

"There are no routine statistical questions, only questionable statistical routines."

—Sir David Roxbee Cox[1]

Contents

Previous chapters presented and discussed the fundamentals for the pedestrian traffic estimation, specifics of episodic movement data from Bluetooth tracking data and details on our two specific complementary methods for pedestrian traffic estimation incorporating movement patterns. Furthermore, we introduced a system for pedestrian

[1] British statistician, born 1924

mobility analysis. In order to confirm the strength of our contributions we applied and evaluated the previously described methods for pedestrian quantity estimation in different real-world industrial scenarios.

In this chapter we will present the applications to the industrial real-world scenarios: *Billboard Location Evaluation* for Swiss train stations, *Visitor Monitoring* at the zoo of Duisburg (Germany) and *Event Monitoring* during a soccer event at Stade des Costières, Nîmes (France).

6.1 Introduction

This thesis focuses on pedestrian quantity estimation methods based on episodic movement data. We developed a *System for Pedestrian Mobility Analysis*, presented in Chapter 5. We moreover applied our contributed methods and system in real-world industrial scenarios. The real world applications we apply our methods to, cover the following real world scenarios:

- *Billboard Location Evaluation* for Swiss train stations: pedestrian quantity estimation in 27 major train stations,
- *Visitor Monitoring* at the zoo of Duisburg: quantity estimation, sensor placement and visitor route choice monitoring and
- *Event Monitoring* during a soccer event at Stade des Costières, Nîmes (France): path selection and quantity estimation in spatio-temporal dimension.

Each of the applications requires different traffic modelling, i.e. quantity estimation based on either *moves* or *visits*. The posed questions in each of the applications trigger our method choices. Thus, besides pedestrian quantity estimation, we provide a comprehensive analysis of the recorded data sets.

In [Stange *et al.* 2011] we introduced a workflow for the application of the pedestrian monitoring system (and refined it in [Liebig *et al.* 2013]). This is an adaption of the general knowledge discovery workflow [Fayyad *et al.* 1996]. This comprises the following steps:

1. The field study phase. This contains survey design and data collection steps.
2. The visual analysis phase contains the data preparation, aggregation and visual analysis steps.
3. The knowledge discovery phase contains the pedestrian analysis step.

The remainder of this chapter is structured as follows: we discuss each of the three applications: *Billboard Location Evaluation*, *Visitor Monitoring*, and *Event Monitoring*. In each of the application sections we follow the mentioned workflow and we roughly present: motivation, field study phase, data preprocessing and data analysis as well as results.

6.2 Billboard Location Evaluation

The following application deals with the problem of billboard location evaluation. In contrast to existing works, the considered locations are not outdoor, but inside public train stations. Therefore the usage of GPS recordings (which is common for this task [Pasquier *et al.* 2008]) is unfeasible. We combine in our approach:

- movement pattern heuristics, which are commonly assumed to hold inside train stations (this assumption was justified in [Li *et al.* 2008]),
- traffic networks derived from public available floorplan sketches and
- (manual) quantitative observations of pedestrian traffic.

Therefore, we apply our LSR method (introduced and discussed in Chapter 3) for the 27 largest Swiss train stations. The study was conducted on behalf of [Swiss Poster Research Plus 2010] who supported us with data collection. In result, arbitrary locations can be evaluated based on their traffic frequency. This output is industrially used for billboard pricing and billboard location planning.

This section is structured as follows. We firstly motivate the importance of billboard location evaluation focussing on indoor locations. Next, we expose the field study phase with details regarding data preparation, i.e. selection and placement of sensors and methods for data acquisition. The actual approach is explained in the next section, *Flow estimation for billboard location evaluation*. Finally we conclude this section with a summary of the billboard location evaluation application.

6.2.1 Motivation

Outdoor advertisement is one of the oldest advertising media and plays an important role in the advertisement industry. In 2008 the turnover was 684 million CHF (about 460 million Euro) in Switzerland and 805 million Euro in Germany [Fachverband Außenwerbung e.V. 2009, Stiftung Werbestatistik Schweiz 2009].

In recent years this advertising market has changed rapidly. The change is predominately caused by two factors, namely the competition with other advertising media and the emergence of digital media. In this application we will focus on a problem that has not been sufficiently addressed by previous methods and that can be extended to dynamic performance measurement in the future. These are performance measurements for billboards that are placed indoors, in buildings. The challenge in this case results from the fact that due to signal loss caused by the building, the GPS trajectories (as used for outdoor billboard location evaluation [Pasquier *et al.* 2008]) just describe which persons enter the building; inside the building itself valid GPS positions are rare and generally not available. Thus, we do not know which person has contact with a particular poster or indoor campaign. In order to overcome the lack of GPS signal we use Bluetooth tracking technologies or empirical measurements at few positions (due to budget constraints) in concert with our previously described traffic estimation method to evaluate the popularity of all corridors (i.e. poster location candidates) inside the building.

We successfully applied this pedestrian quantity estimation method to the data collected in 27 major Swiss train stations. Within this section we illustrate the process for one example: Zurich central station. It highlights the advantage of the proposed method combining the calculated model with existing outdoor poster campaign evaluation approaches.

6.2.2 Field Study Phase

The data we base our analysis on are (1) floor plan image (2) empirical recordings of pedestrian quantities at pre-selected locations and (3) a heuristic on pedestrian route choice. Before the previously presented monitoring system (compare Chapter 5) can be applied, a required pre-processing step is the traffic network construction based on the floor plan image.

The representation of the walkable area by networks is reasonable, since pedestrians may only enter and exit the area via dedicated exits (entry and exit points). Edges of the so-called *traffic network* (compare Definition 7 in Section 3.2) represent the corridors and paths, whereas junctions are represented by vertices.

Algorithms for traffic network construction with a given floor plan image are so-called *triangulation methods*. Triangulation methods are closely related to tessellations as they may be mapped to each other: every vertex of a triangulation becomes the handle of a surrounding polygon. A commonly applied automatic method for traffic network construction is the Delaunay triangulation [Delone 1934], its tessellation counterpart is the Voronoi tessellation [Dirichlet 1850, Voronoï 1908] (introduced in Section 2.2.4.1). The Delaunay triangulation algorithm constructs a mesh of adjacent triangles among a given set of vertices. The process which automatically derives a traffic network from a floor plan is described in [Demyen & Buro 2006]. Briefly summarized, (1) vertices are drawn uniformly at random from the walkable area of the floor plan image, (2) the Delaunay triangulation among these vertices returns the traffic network.

Another possible method for traffic network construction from a given floor plan is the manual creation of vertices at junctions and connecting edges among the corridors. This method does not depend on exact geometric representation but may also handle floor plan sketches, if they provide the topology of the walkable area. An example of the traffic network for the hereby considered Zurich central station is depicted in Figure 3.2. After the network's structure is obtained, exits are marked manually.

After performing a pre-study we concluded that counting the number of people manually at several positions (using a smart phone application for data entry) is the most cost-efficient method for data collection. As noted in Chapter 2, using video cameras was not feasible because of privacy constraints. To decrease the influence of the day of week on the measurements, we repeated the measurements at three different days. As the number of "sensors" (countings) is limited, we had to select locations for counting in advance using the traffic network of the train stations. Therefore we located sensors at the most important junctions and stairways. Figure 6.1 depicts the measured edges at Zurich central station [Liebig *et al.* 2010].

To assist manual counting and to simplify post-processing of measurements, we developed a smart phone application (Figure 6.1) which records clicks of the survey-

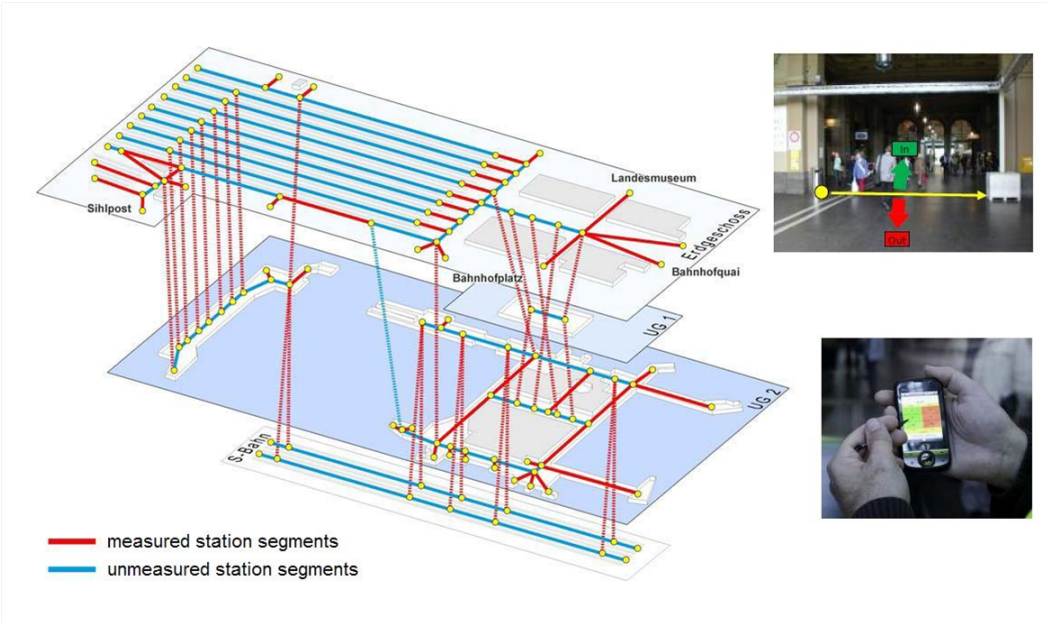

Figure 6.1: Sensor location and smart phone application for manual counting (image credits [Swiss Poster Research Plus 2010]).

ing person - each click represents the number of pedestrians passing by in a specified direction - along with its time-stamp. This enables easy integration of the measured data in the monitoring system as part of the *Sensor Layer* (presented in Section 5.4.1). Thus, we know how many people passed at which time into which direction.

6.2.2.1 Data Preparation

After the previously described *field study phase* and data collection it is necessary to prepare the data. This data preparation enables comparison of the empirical raw data of pedestrian quantities at the measurement locations. After fetching the measurements, the counted quantities are weighted and aggregated according to space (the edge, i.e. the corridor, of the traffic network) and time interval (i.e. the day they were taken). As a result, every measured location in the train station has associated with it a number of pedestrians that may be compared against any other measured location. This is important for ranking locations or segments within the building, which is a first feature of the hereby introduced analysis.

Since the traffic frequencies are only recorded at sensor locations the estimation of pedestrian quantities is necessary for unobserved locations. Next section addresses this task.

6.2.3 Flow Estimation for Billboard Location Evaluation

For segments where empirical measurements have been taken, the quantities are known. Triggered by the *Query Layer*, our task is now to estimate them for the unobserved segments, and to build a pedestrian indoor movement model that is useful for

poster and campaign evaluation within the selected time period.

For billboard evaluation not just traffic quantities for single locations, but also their correlation is of high interest. This is the question for plausible trajectories which are consistent to the model assumptions and also fit to the estimated frequencies. Such a set of trajectories denotes how many people being at one particular location also pass any other. Therefore this could be utilized for the estimation of how many different people have contact with a billboard campaign (consisting of multiple posters within one train station) in the considered time interval.

Our previously presented method (Section 3.6) addresses this challenge in a regression approach. In a first step, we enumerate all plausible routes through the building and collect them in a route set. For example, at the main station in Zurich, there are about 380,000 conceivable routes. Non-plausible routes are eliminated, among them circular routes. Afterwards, we assign frequencies to each route, based on the measurements and the movement pattern heuristic. We assume that pedestrians prefer for a walk from a particular entrance to another exit the path with the minimal detour, therefore the detour of the path is calculated and becomes a feature of the path. This detour of a path is the characteristic function of the movement pattern heuristic introduced in Definition 11.

The sensor readings, traffic network and movement pattern heuristic (in combination with the enumerated set of plausible paths) are passed to the previously introduced regression algorithm (Section 3.6) which estimates frequencies for unobserved locations and weights for the paths. The measurements serve as frequency targets in this regression process. The purpose of this assignment procedure is to find an optimal combination of routes that fulfills all frequency targets [Liebig *et al.* 2010]. For convenience, this process is sketched in Figure 6.2.

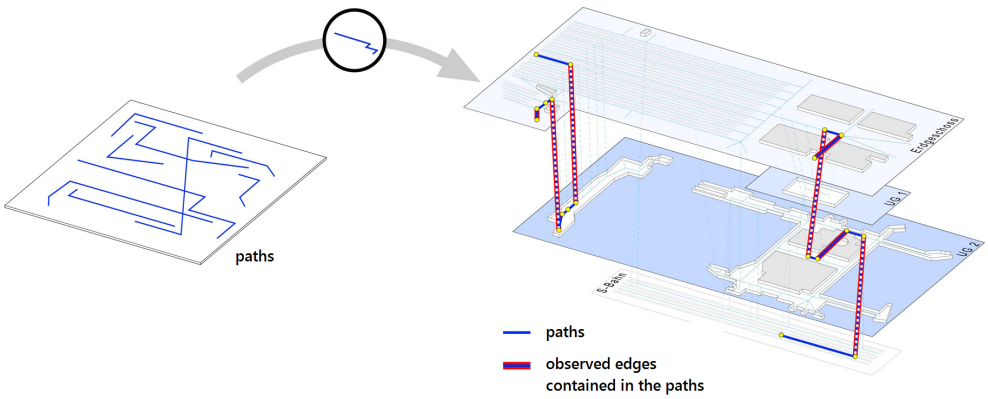

Figure 6.2: Method of Traffic Quantity Estimation (image credits [Swiss Poster Research Plus 2010]).

As a result we obtain for every modelled train station (1) a set of routes crossing that station and (2) the number of people walking on each route. Figure 6.3 gives an example for Zurich central train station [Liebig *et al.* 2010]. With this information we

are able to calculate quantities for each edge in the station by summing over route quantities, no matter whether the particular edge has been measured empirically or not. This yields the pedestrian movement model based on empirical measurements we aim for. It enables us to denote pedestrian quantities at any location and gives trajectories also at unobserved segments.

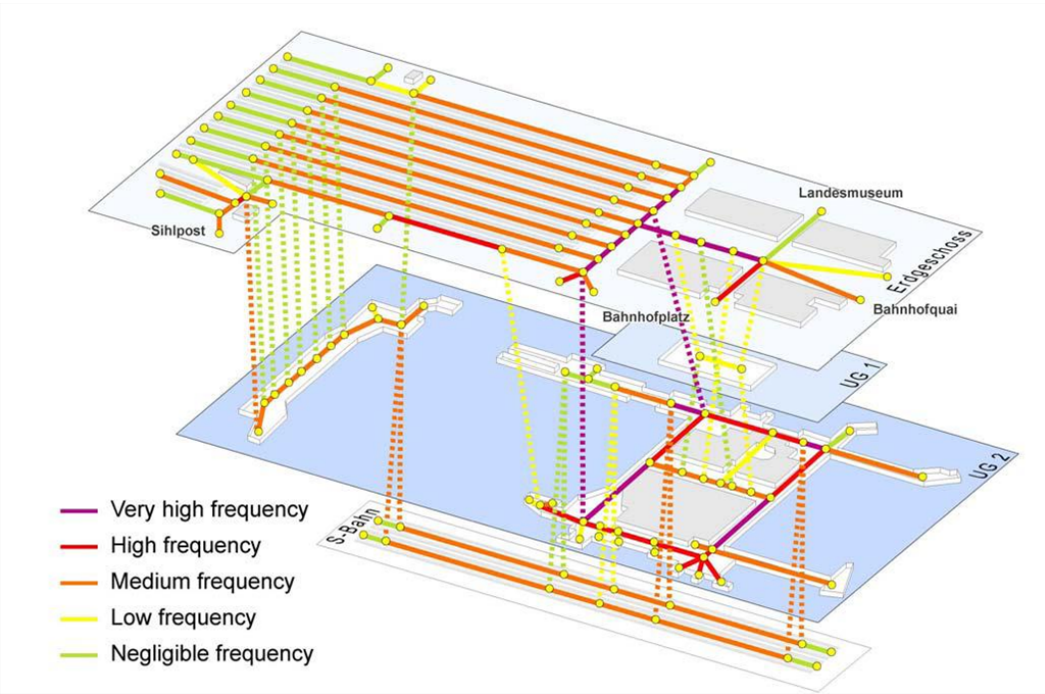

Figure 6.3: Result of Traffic Quantity Estimation (image credits [Swiss Poster Research Plus 2010]).

6.2.4 Integration with GPS Surveys

In order to apply the frequency estimation results for the evaluation of mixed indoor-outdoor poster campaigns, we need to integrate them with GPS mobility data as mentioned in the introduction. This means that we have to re-use the output of the *Analysis Layer* and assign for each GPS person who enters a railway station a corresponding route through the station[2]. To achieve this goal, we need to perform three subsequent steps:

- (1) visit identification,
- (2) route assignment and
- (3) performance evaluation.

[2]**published with a major contribution of the author in:**
T. Liebig. *Trajectory Regression Model for Indoor Pedestrian Flow Analysis on Billboard Evaluation*. In Proc. of the Third International Conference on Applied Operation Research - ICAOR'11, pages 289–300. Tadbir Operational Research Group Ltd., 2011.

In the first step we identify all test persons within the GPS sample visiting a major Swiss train station. Based on GPS trajectories of over 10,000 test persons recorded over a period of one week per person, we isolate all tracks in the vicinity of a train station using buffers and the spatial join operation "intersect" (this operation filters the GPS points which are inside the buffer zone, and belong to the visit, Definition 3, to the station, compare Section 2.2.4.1). As GPS signals may be noisy, we apply an individually sized buffer to each of the train station geometries, reflecting its specific local setting. The resulting candidate set, however, contains not only potential rail travellers but also regular pedestrians, car drivers and passengers passing by the station without entering it. We therefore apply a complex multi-level filtering process which identifies the visitors of a train station using, for instance, speed curves, the course of movement and time spent inside the geometric extension of the train station. Knowing all visits to a railway station completes step one.

Step two is the assignment of each visit to one of the routes underlying the pedestrian movement model. The challenge of this task is to find an optimal distribution of personalized routes given the route frequencies. We do this iteratively by drawing routes from the route set and considering the projected weight and sociodemographic information of each test person being assigned to that particular route. At the end of this step each GPS trajectory containing a visit to a train station, as identified in the previous step, has been assigned a route through the corresponding train station as depicted in Figure 6.4.

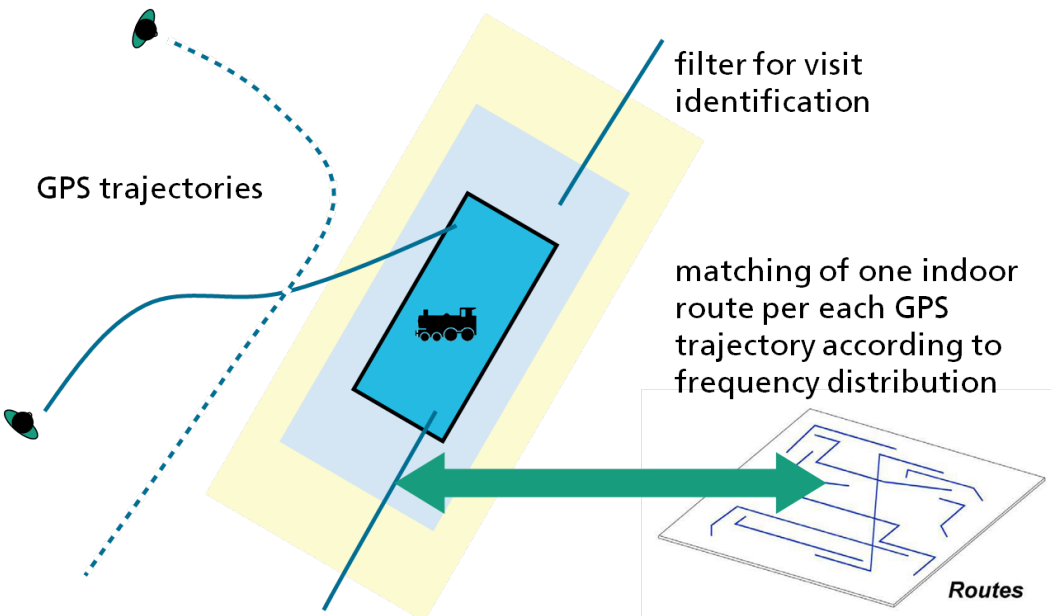

Figure 6.4: Integration of indoor/outdoor billboard location evaluation scenario. Filters on outdoor GPS trajectories are applied to identify train station visits. Every visiting GPS trajectory is matched to an indoor route according to their frequency distribution.

Finally in step three we weight poster contacts and calculate performance measures of mixed in- and outdoor campaigns. Similar to the performance evaluation of

outdoor posters, we consider individual visibility criteria at each poster site. Routes passing the visibility area are weighted according to the contact quality, depending e.g. on the viewing direction or clustering of panels. Given weighted contacts for each indoor poster and visiting person, we can estimate total contacts and reach of a mixed indoor-outdoor campaign using the same algorithmic background as for outdoor campaigns. The selection of a campaign and of a target audience depends on all relevant (indoor and outdoor) poster contacts and the application of Kaplan-Meier compensates for missing measurement days in the GPS data as described in [Pasquier *et al.* 2008].

6.2.5 Summary

In this industrial billboard location evaluation application scenario we developed a workflow that allows performance measurements for billboards that are placed indoors. We focused on 2,600 poster sites in railway stations as those are being seen as one of the most valuable over all. The challenge results from GPS signal loss inside buildings. Our proposed pedestrian monitoring system includes the development of a pedestrian model based on empirical data. This mobility information has been integrated with existing GPS mobility data, allowing to infer reach values and weighted contacts. We applied our approach to 27 major Swiss train stations. Although we showed how to implement a general movement model within the train station which is used for poster campaign evaluation, we do not model time, so far. Our indoor model is static for the considered time period. Whereas this is already sufficient for current billboard evaluation, continuous triggering of the analysis in a real-time scenario could be used for pervasive advertisement evaluation.

6.3 Visitor Monitoring

This real-world application scenario deals with the monitoring of visitors to a zoological garden and highlights the usage of Bluetooth sensors not only indoor, but outdoor as well. We analyse the obtained tracking data in order to obtain information on stay times and route choice preferences. These preliminary studies are a joint work published in [Ellersiek *et al.* 2012]. Afterwards we apply our quantity estimation method GPR presented in Section 3.7 incorporating the movement patterns recorded by use of the Bluetooth scanners.

In this application we deal with *Episodic Movement Data* (introduced in Section 2.2.4) and we apply methods for analysis of this mobility data type for the considered real-world application. Besides quantity estimation, an important new problem addressed in this application, is automatic sensor placement, for which we provide a solution as well.

The application is exposed in the following section as follows. Firstly, we motivate the analysis and highlight the questions we are going to tackle in this application. Secondly, we present the *field study phase* (data collection performed in [Ellersiek 2011]). In the following sections we discuss the *data preparation* consisting of representativeness of the tracking technology and evaluation of spatio-temporal de-

pendencies [Ellersiek *et al.* 2012]. The actual focus and main author's contribution, i.e. traffic quantity estimation for visitor monitoring, as well as its results is tackled in the section *Traffic Quantity Estimation for Visitor Monitoring*[3] (Section 6.3.6) followed by the results for automatic sensor placement for this scenario. Finally, the section closes with a short summary of the application and analysis.

6.3.1 Motivation

Monitoring the visitor behaviour of a museum or a zoological garden reveals insights on preferences and enables improvements of the signage, the infrastructure or in case of a museum also on the exhibition itself [Noschka-Roos 2003, Scholz 2009]. It has become a common practice to analyse the visitor behaviour for these public exhibitions [Hase 2011]. In the application scenario at-hand we perform the evaluation of visitor movement behaviour based on Bluetooth tracking (Chapter 4) data in the zoological garden of Duisburg, Germany.

The analysis tackles two major problems: (1) location-based analysis and (2) trajectory-based analysis. The first one focuses on locations which were equipped with Bluetooth sensors and addresses following questions:

- Which attractions are most popular?
- How long in average does a person visit an attraction?
- Do spatio-temporal dependencies exist among the visitor quantities of the attractions?

The aim of these location-based analyses is the extraction of the attractiveness of the locations to get an overview on which locations are frequently visited and which ones are less frequently visited. Hereby it is studied how many people visit a location and how long their average stay time at the location is. Furthermore, the recorded pedestrian frequencies of particular locations should be compared to the frequencies of other attractions in order to detect whether there exist causal dependencies among the visitor presences, e.g., fluctuations of visitor numbers which could increase at one location and simultaneously decrease at another one.

Besides the location-based analyses at various locations among the zoo, we study the spatio-temporal dynamic of visitor frequencies within the closed site. In this context Bluetooth tracking provides its benefits. It does not only support location-based analysis as frequency studies (compare previous section), but due to the recording of unique identifiers for each recorded Bluetooth-enabled mobile phone or intercom it also enables detection and analysis of mobility patterns among multiple sensor-footprints. This second class of analyses, the analysis of movements, focuses on:

- Are there regular mobility trends which are repeatedly monitored?
- Which paths are most frequently used?

[3] **published with a major contribution of the author in:**
T. Liebig, Z. Xu, M. May and S. Wrobel. *Pedestrian Quantity Estimation with Trajectory Patterns*. In Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases ECML PKDD 2012, Part II, LNCS 7524, pages 629–643. Springer, 2012

In order to answer the questions the study was conducted in the steps recalled from [Liebig *et al.* 2013] in Section 6.1. These steps *field study phase*, *data preprocessing* and *data analysis* as well as the presentation of *analysis results* are the structure of the next sections.

6.3.2 Field Study Phase

The project from which this application is derived lasted 19 days during July and August 2011. For the location-based analyses sensors were situated next to the animal compounds as well as close to gastronomic facilities and at both zoo entries and exits. Furthermore, sensors were placed at most important, i.e. highly visited junctions to enable path analyses in the second step. The sensor locations are depicted in Figure 6.5. As mentioned in Chapter 4 two different types of Bluetooth antennas were integrated in the sensors with a range of either 20 meter or 100 meter. The Bluetooth beacons placed inside buildings were equipped with 20 meter antennas instead of the 100 meter antennas to prevent the sensors from monitoring also people passing by next to the building and bound their footprint just to the inner building. We checked proper behaviour after installation. The sensors equipped with 20 meter antennas were placed in the buildings, whereas the ones with 100 meter antennas were placed outside. More detailed information on the field sensor positions can be found in [Ellersiek *et al.* 2012].

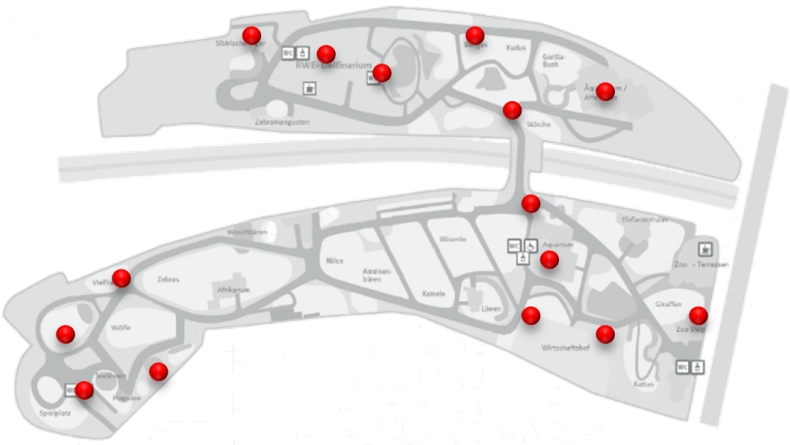

Figure 6.5: Locations of the Bluetooth Scanners (red dots) at the Zoo of Duisburg [Liebig *et al.* 2012b].

During the field study phase approximately 500,000 data points were recorded, which were produced by about 7,000 unique Bluetooth devices with enabled Bluetooth visibility mode. But not all of these data point can be used for the analyses as not all devices have been recorded during the whole visit in the zoo, but just partially due to several reasons (compare section on *Episodic Movement Data*, Section 2.2.4). In this study all recorded Bluetooth devices that stayed more than 30 minutes and less than 10 hours in the zoo, are considered. The average stay times in hours for the days during the study are depicted in Figure 6.6.

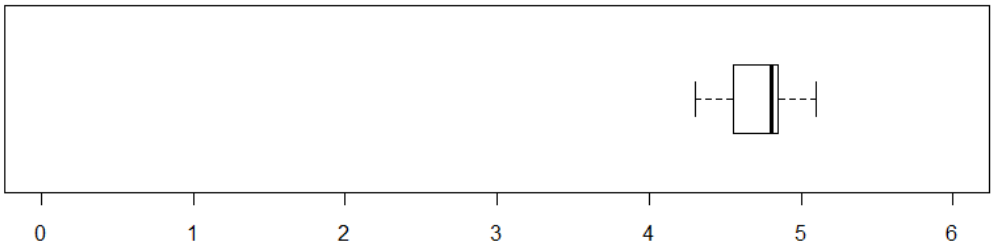

Figure 6.6: Condensed representation of the average daily stay times of the visitors at the zoo of Duisburg in hours (the x-axis shows the number of hours). Compare Appendix B for an introduction to box-whisker plots.

6.3.3 Representativeness

The representativeness of the Bluetooth tracking technology, evaluated in practical applications in the *Event Monitoring* application scenario, has been discussed in Chapter 4. However, since this is a key topic for the real-world applications based on Bluetooth sensor data, we review the representativeness results in this particular application scenario.

For estimation of the representativeness of the recorded data sample, the recorded number of people was compared to the total number (estimation based on number of sold tickets) of visitors during this period. The comparison reveals that the pedestrian movement contained in the collected Bluetooth tracking dataset represents about 6% of the present visitors [Ellersiek *et al.* 2012]. Thus, the presented empirical analysis confirmed temporal representativeness, which remains almost constant for the different days.

6.3.4 Location Based Analyses

For analysis of the location-based questions, the aim is to rank the attractiveness of the compounds based on number of visitors and stay times and to identify spatio-temporal dependencies among the different locations in the zoo, tackled in [Ellersiek 2011]. Therefore, unique visitor identifiers are considered. The average stay time and the total number of the recorded Bluetooth devices (respectively the visitors) to the individual compounds can be aggregated as depicted in Table 6.1. The table reveals that the dolphinarium has longest stay times and the aquarium the most visitors, whereas the koala house appears to be less attractive to the visitors.

Furthermore, the temporal dynamics of visitor quantities was aggregated on time-slices of 30 minutes and afterwards compared for selected attractions [Ellersiek *et al.* 2012]. In this context four compounds were analysed: the dolphinarium, the equatorium, the aquarium and the koala compound. Five days of the dataset were considered for analyses; at these days the four dolphin shows happened to be at the same time. Visitor numbers were aggregated in 30 minute intervals and averaged for the 5 days (see Figure 6.7).

compound	number of BT-devices	average stay time (in min)
dolphinarium	1,923	46
equatorium	2,198	14
aquarium	2,649	11
koala house	1,483	7

Table 6.1: Number of Bluetooth devices and average stay times (in minutes) at the compounds in the zoo of Duisburg

Whereas the visitors in the dolphinarium house rise in advance of the dolphin shows, the visitor quantities at the aquarium and the equatorium decrease. Similarly, the visitor number at the latter rises after the end of the dolphin show. A dependency among visitors at the koala house and the dolphinarium could not be recognized, due to their large spatial distance.

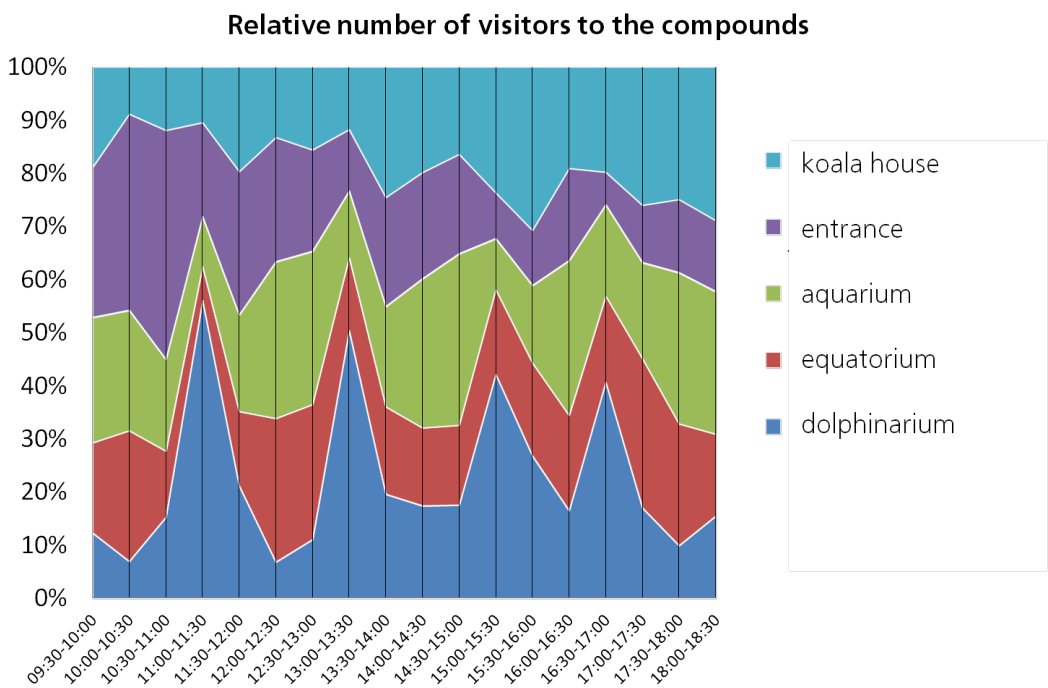

Figure 6.7: Spatio-temporal dependencies of visitor counts at the Zoo of Duisburg.

This assumption on spatial dependencies rises the question for movement preferences and introduces the trajectory analyses, we performed next.

6.3.5 Trajectory Analyses

For the path-based analyses continuous movement recordings of visitors among multiple sensor locations are interesting. Though, these episodically recorded trajectories can not be visualized by line interpolations but by flows, they consist of individual sequences of multiple sensors (compare Chapter 4), because visitors pass multiple lo-

cations equipped with sensors during their visit [Stange *et al.* 2011]. Thus, these location sequences consist of a temporal order of the location visits and are very versatile, because every visitor chooses his individual route through the zoo. The aim of the path-based analysis is to identify location-sequences among the dataset, which occur since multiple people chose the same path through the zoo. Thus, the dataset needs to be preprocessed accordingly. The collected episodic dataset is not temporally sparse, but contains multiple consecutive data points for a particular mobile phone per location, because the Bluetooth scanner records a data point for every scan. Based on the usage of three Bluetooth antennas, the scan processes are performed about every three seconds (compare Section 4.2.1 on the details of the hereby applied Blutooth sensor technology). The recorded sequences of individually passed locations contains duplicates which are removed before further processing (for details, see [Ellersiek 2011, Ellersiek *et al.* 2012]). Thus, the raw sequences are reduced to sequences of *visits* or *moves*. As proposed in Chapters 1 and 4 the number of pairwise *moves* among two locations can be aggregated (for all persons) and depicted on a *flow map* [Kraak & Ormeling 2003], see Figure 6.8. The dataset used for creation of Figure 6.8 comprises 931 visitors. This is the set of persons which have been recorded at the entry and the exit of their visit and started their journey at the main entrance (position number 1 in the Figure 6.8). The arrows depict sequences of their movements. In brackets the number of persons and the percentage is depicted [Ellersiek *et al.* 2012].

Considered are just visitors, who start their visit at the main gate of the zoological garden. The reason is that more similarities of the visit sequences are contained in the beginning and the end of a visit sequence. As presented in [Hägerstrand 1970] the locations and movements become more diverse during the movement in the zoo, because chosen routes become more individual, and just the fixed start and end point (situated at the main entrance) constrain the *space time prism* (see Section 2.2.1 on time geography) of the individual movement. Though Figure 6.8 represents movement behaviour just from the beginning of the sequence, the visitor trajectories do not end at the place where no arrows are depicted, but at the main exit.

After starting the visit at the main entrance, it is noticeable that most of the visitors walk directly to the aquarium, location number three. The compound of the lemures is not attended, and the walk is continued in the direction of the dolphinarium, position 11, in the eastern part of the area. While just a few persons continue their visit among the western half of the area, Figure 6.8 also shows that many people are heading to the dolphinarium, position 11, using the shortest path. Probably this results from the fixed time schedule of the dolphin shows. For more details on this statistical analysis of the dataset, we refer to [Ellersiek *et al.* 2012].

6.3.6 Traffic Quantity Estimation for Visitor Monitoring

In order to estimate visitor numbers at unobserved sensor locations, we apply the proposed method (Section 3.7) to the collected dataset of visitor movement in the zoo of Duisburg (Germany). The dataset consists of episodic movement data [Andrienko *et al.* 2012] and was collected with a mesh of 15 Bluetooth scanners within the period of 7 days (07/26/11–08/02/11).

In order to perform the tests, the traffic network is build from the sensor positions.

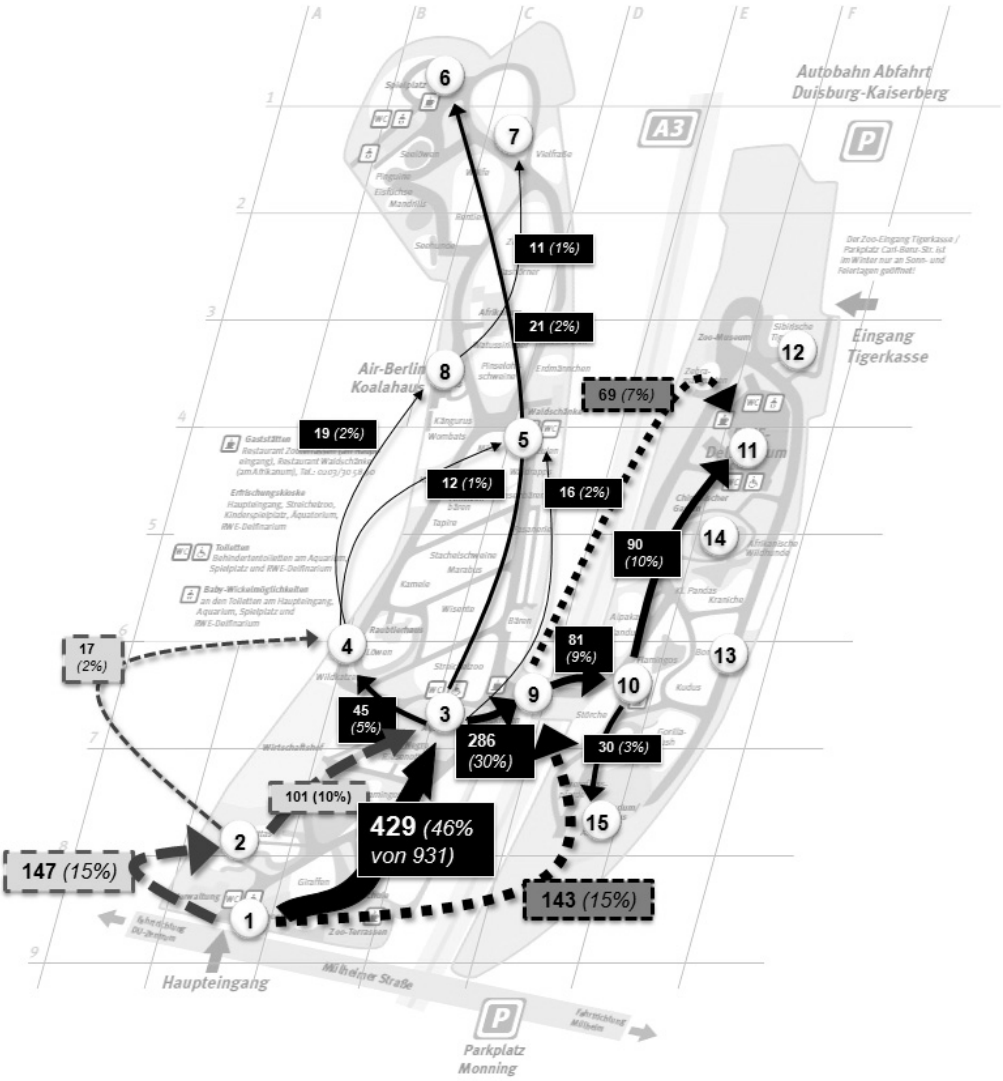

Figure 6.8: Visitor flows at the Zoo of Duisburg starting at the main entrance [Ellersiek *et al.* 2012].

Each sensor becomes a vertex of the traffic network. To achieve ground truth for the traffic volume prediction, temporal aggregates of recorded transitions between sensors, as proposed in [Andrienko *et al.* 2012], are scaled by the Bluetooth representativeness (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 (i.e. edges) 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 required movement pattern input to the *movement pattern kernel* (compare Section 3.7). Similarly to the previously synthetically generated data, the real world experiments are conducted with different percentages of measured edges (10% to 50%). These measurement edges are

chosen uniformly at random 100 times for each test dataset.

Similar to Section 3.7 we compare our introduced quantity estimation method to state-of-the-art approaches (Spatial kNN, and Gaussian Process Regression with commonly used kernel functions) for each of the 5 datasets. The accuracy of the application to real-world data is also measured in mean absolute error MAE, which became a common measure of traffic volume estimations [Zhao & Park 2004, Gong & Wang 2002, Neumann *et al.* 2009].

Results of the experiment are depicted in Figure 6.9, the proposed method (represented by the box-whisker plot in the middle, called PATT) achieved the best prediction performance (measured in mean absolute error - MAE) for the pedestrian quantity estimation problem in comparison to other state-of-the-art methods. The tabular view on the condensed MAE distribution can be found in Table C.2, Appendix C. Especially for the first dataset which represents a sparsely monitored zoo traffic network, the results outperform existing methods. The incorporating of expert knowledge on movement preferences allows for the model to well capture the dependencies of traffic at different edges and, moreover, to improve prediction accuracy.

The hereby described real-world application was also discussed in Section 5.5 in the context of the software integration. There we described design and development of an industrial applicable software system for this analysis. In this software system, the traffic network was not generated by the sensor locations but a predefined network was applied and map matching was required for matching of the sensor positions to the network. The computation, validated in this section, and described in Section 3.7 was integrated in a web-based user interface which provides map representations of the results (Figure 5.4).

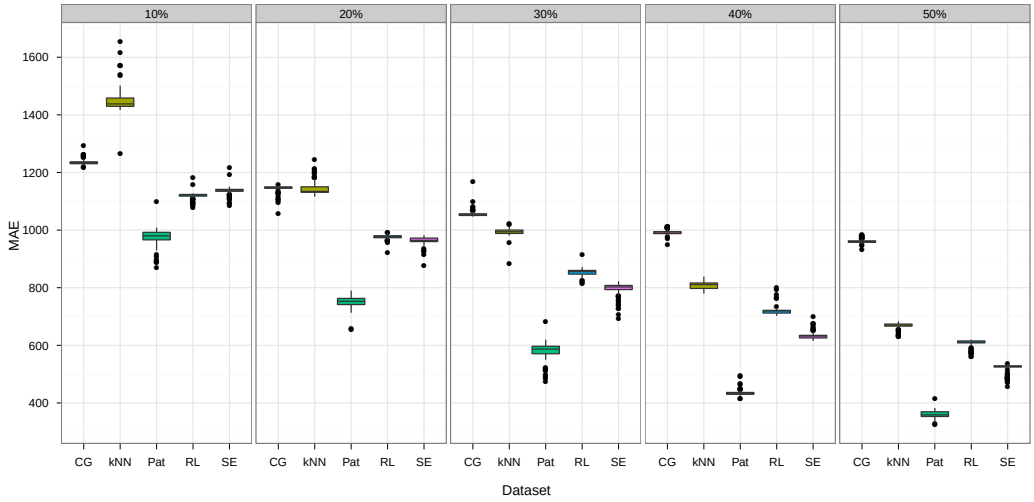

Figure 6.9: Quantity estimation performance at the zoo of Duisburg [Liebig *et al.* 2012b]. Performance is measured by mean absolute error (MAE) at settings with different ratios of monitored edges (10 to 50 percent). The five methods are: GPR with diffusion kernel (Diff), spatial k-nearest neighbour (S-kNN), GPR with trajectory pattern kernel (Patt), GPR with regularized Laplacian (RL) and GPR with squared exponential kernel (SE). Compare Appendix B for an introduction to box-whisker plots and Table C.2 for a tabular representation of the depicted values.

6.3.7 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 quantity estimation method, namely the movement pattern kernel, we perform the sensor placement procedure on the zoo of Duisburg data. Afterwards, pedestrian quantity estimation based on resulting sensor placement (instead of random placement as done in previous section) is carried out and performance is measured with MAE (mean absolute error). The computational process is explained in Section 3.7.1.

The experiments are conducted similarly to previous section. Again, we consider five different cases, depending on the percentage of edges which are observed by sensors (ranging from 10% to 50%). For each of these ratios, we compute the automatic sensor placement and hand the readings of the resulting positions to the movement pattern kernel, which estimates the quantity at all remaining (unobserved) edges. The error is again measured in MAE, mean absolute error.

As result, the red horizontal line in Figure 6.10 depicts the sensor placement performances in comparison to previous random placement (tabular details are Appendix C, Table C.2). 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 movement patterns.

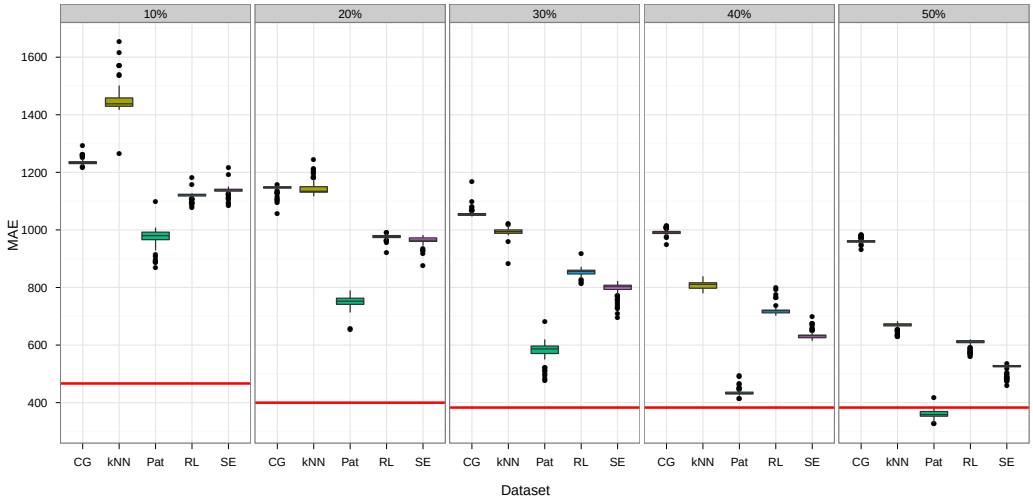

Figure 6.10: Sensor placement performance at the zoo of Duisburg [Liebig *et al.* 2012b]. Performance is measured by mean absolute error (MAE) at settings with different ratios of monitored edges (10 to 50 percent). The five methods for comparison are: GPR with diffusion kernel (Diff), spatial k-nearest neighbour (S-kNN), GPR with trajectory pattern kernel (Patt), GPR with regularized Laplacian (RL) and GPR with squared exponential kernel (SE). Compare Appendix B for an introduction to box-whisker plots and Table C.2 for a tabular representation of the depicted values.

6.3.8 Summary

This section focussed on automatic, quantitative visitor monitoring for a zoological garden. The analysis of visitor behaviour became a major topic for museums and zoos. We conducted our experiments using 15 Bluetooth sensors for 19 days in July and August 2011. The data analysis was performed using the previously introduced workflow. These steps *field study phase*, *data preprocessing* and *data analysis* as well as the presentation of *analysis results* were the structure of the last sections.

The tackled questions divided in two classes, (1) location based questions and (2) trajectory based questions. Therefore we addressed first:

- Which attractions are most popular?
- How long is a visit of a person to an attraction?
- Do spatio-temporal dependencies exist among the visitor quantities of the attractions?

and afterwards:

- Are there regular mobility trends which are repeatedly monitored?

- Which paths are most frequently used?

For the last question, we presented in this section the application of previously introduced methods for visitor monitoring in a zoological garden. For this real-world application we used a nonparametric Bayesian method to tackle the pedestrian quantity estimation problem which explores the expert knowledge of movement patterns. We validated our proposed method on the real-world dataset collected with the help of Bluetooth tracking technology at the zoo of Duisburg. Furthermore, we addressed the question for sensor placement in the given industrial scenario with the trajectory based graph kernel. The empirical analysis demonstrated that incorporating movement patterns can largely improve the traffic prediction accuracy in comparison to other state-of-the-art methods. This application has shown that our work also provides an efficient and applicable solution to pedestrian volume estimation in industrial real world scenarios.

6.4 Event Monitoring

This section presents a use case at a soccer stadium. On behalf of the European Emergency Support System Project [ESS 2010] (described in previous Chapter 5), we placed 17 Bluetooth Scanners at various locations of the stadium in order to track visitor movements. Aim of the project was the establishment of a information and control system which supports local forces in event monitoring, crisis management and risk analysis. In contrast to existing C4i (*Command, Control, Communications, Computers, and Intelligence*) systems [Ebbutt 2011] the system integrates expert analysis modules and rich communication capabilities. Additional advantage is that all the used interfaces and protocols are open and well documented, which enables further development of components without dependency on single software companies. Various sensor technologies should integrate seamlessly in the system and the system should not depend on a particular communication media (GSM, WLAN). These goals were achieved, by usage of open communication protocols and inclusion of external expert systems. The external services are provided as RESTful web services (this REpresentation State Transfer architecture was introduced in Chapter 5, an introduction is also given in [Richardson & Ruby 2007]).

The historical data analysis component we provided (see Section 5.5.3) hands in the spatio-temporally aggregated visitor quantities (presence situations) during previous events to the local forces. The analyses we performed with the recorded dataset are:

- clustering analysis of visitor presence and flow
- creation of a probabilistic model of inner trajectory correlations
- pedestrian quantity estimation with movement pattern and thereby analysis on possibilities to omit frequency sensors

The analysis was conducted in the steps introduced in Section 6.1 [Liebig *et al.* 2013]. These steps *field study phase*, *data preprocessing* and *data analysis* as well as the presentation of *analysis results* structure the next sections.

We structure the section as follows. Firstly, we shortly expose the motivating facts for this application. Next, the field study phase is described with details on the sensor placement. In the next section we tackle the data description and visual analysis on the collected data for this scenario. Next sections deal with the analysis of presences and respectively analysis of flows, continued by the visual analysis of the inner trajectory correlations. The actual approach is discussed in the last section for this application, *Traffic Quantity Estimation for Event Monitoring*.

6.4.1 Motivation

Major public events as concerts and soccer matches which attract thousands or millions of visitors are on one hand a great chance for street marketers and advertisement companies but on the other hand also a growing financial risk for the organizers and a safety hazard for the guests themselves, due to huge expenses and high visitor densities. Whereas one trend is to build larger infrastructures (concert halls, stadiums) another trend is the growing visitor number at events. In the last years, this hazardous development led to devastating disasters (e.g. at Loveparade festival in 2010). Thus, visitor monitoring in complex facilities became an important subject. But understanding the movement behaviour, identification of attractors and distractors, determination of waiting times, as well as localization of congestions and bottlenecks gives also insights on visitor preferences and motivations at a particular site or event. Knowing such detailed information on indoor pedestrian behaviour gives also a location based performance indicator for different locations inside the building. Various locations and attractions can be ranked by their popularity, safety or frequency. Recently evolved Bluetooth tracking (Chapter 4) became the state-of-the-art method for combined indoor outdoor monitoring of pedestrian movement [Andrienko *et al.* 2012, Hagemann & Weinzerl 2008, Leitinger *et al.* 2010, Liebig & Kemloh Wagoum 2012, Stange *et al.* 2011, Versichele *et al.* 2012b]. Understanding movement behaviour or identification of attractors and distractors gives insights on visitor preferences and motivations during a particular event. Various locations and attractions can be ranked by their popularity, safety or frequency.

6.4.2 Field Study Phase

As discussed in Section 4.2, we equipped our sensors with multiple Bluetooth antennas which search simultaneously for visible Bluetooth devices within the sensor footprint. Thus, a complete scan of the frequency band is accelerated and moving people are more likely to become detected while they are in the sensor's footprint. Each time a Bluetooth device (e.g. smartphone or intercom) is recognized by the sensor a data entry is stored in a file. This log-entry consists of time-stamp, sensor identifier, unique hashed device identifier and the signal strength. The need for hashed device identifier results from the fact that Bluetooth sensors collect privacy sensitive data. Every Bluetooth chip is identifiable by its unique Media-Access-Control-address (MAC) [Bluetooth SIG 2004]. Hence, a Bluetooth device (respectively a person) is detectable (and therefore trackable) beyond the spatial-temporal boundaries of an event. Hence, our Bluetooth-scanners save just an anonymized identifier, valid for the time of the mon-

itored event. To scramble the MAC-address, we embed the irreversible SHA-256 encryption algorithm [National Institute of Standards and Technology 2002] with an event specific random seed into the sensor software. Thus, from the very first data recording the tracking of people beyond the event boundaries is unfeasible for the collected dataset.

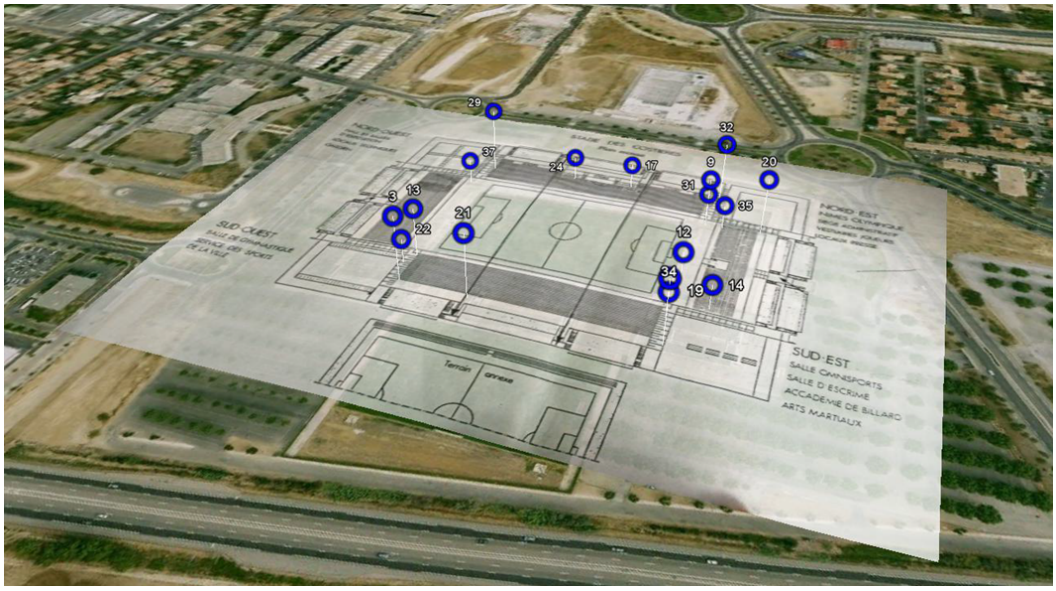

Figure 6.11: Sensor Placement at Stade des Costières Nîmes [Liebig *et al.* 2013].

In the first step, we conduct a *field study phase* in order to locate most suitable places for sensor placement. The location candidates provide power plugs and are at neuralgic indoor places all over the site. The scanners were installed in advance to the match and remained at their locations till the departure of the visitors. Thus, complete trajectories of the visitors are contained in the data. For data collection a mesh of 17 sensors has been deployed among the soccer stadium (Stade des Cosières, Nîmes at France) during a soccer match on 08/05/2011. The three-dimensional sensor placement is depicted in Figure 6.11.

Every single sensor runs asynchronously and the maximum time for each Bluetooth scan may easily exceed the theoretical upper bound [Woodings *et al.* 2001,Bruno & Delmastro 2003]. One reason is the noisy environment during the event. Thus, temporal filtering and spatial aggregation is necessary to get a pure dataset. Duplicate entries are removed from the dataset as well as devices which were only detected once (spot readings). The sensors also record by chance devices of non-interest (e.g. navigation systems of cars or pedestrians passing by at the border of the event area). These artifacts are removed by *vendor filtering* (based on the first three bytes of the MAC address that identify the vendor [IEEE 2002]) as well as spatial-temporal filtering (based on recorded radio signal strength and time-stamp). Finally, arbitrary jumps among sensor locations, resulting from overlapping sensor footprints are also removed. For this step, the spatial distances between sensor footprints, time-stamp and duration of the stays are taken into account to calculate individual position changes per time.

After the data has been purified the dataset contains sequences of positions visited per device enriched with the time-stamp and stay-time duration. Thus, for every sensor location quantity of detected visitors can be determined within a particular time interval by aggregation (compare Section 2.2.4.1 for spatio-temporal aggregation). Additionally, movement patterns and information on the popularity of sensor location transitions remained preserved in the data.

We recorded 47,589 data points from 553 different devices at 17 distinct locations. The average number of distinct visited sensor locations is 4.37, the median number is 2. The recorded movements have an average duration of 3 hours and 25 minutes. In total, about 14 percent of the visitors, 553 of 3,898 (this official visitor number does not contain the people which worked there[4]), have been recorded during the period of the match, thus we expect the dataset to be representative. More detailed analyses and experiments on spatio-temporal representativeness of Bluetooth tracking in a soccer stadium was conducted in Section 4.2.2.

6.4.3 Data Description and Visual Analyses

Episodic movement data consist of records including the following components: object identifier o_k, spatial position p_i, time t, and, possibly, other attributes (compare previous section on *Episodic Movement Data*, Section 2.2.4). The spatial position may be specified directly by spatial (geographic) coordinates $p = (x, y)$ or $p = (x, y, z)$ or by referring to a sensor or location having a fixed position and dimension in space. A chronologically ordered sequence of positions of one moving object can be regarded as an abstract trajectory which is spatially and temporally discontinuous (compare our definition of *Movement Patterns* in Section 3.2, Definition 10).

For temporal aggregation of the data, the time is divided into consecutive intervals of 15 minutes. For spatial aggregation, it is necessary to define a finite set of places visited by the moving objects. For spatial aggregation, we choose the first of the two different cases for spatial aggregation, which we distinguished in Section 4.4.2:

- The object positions in the data are limited to a finite set of predefined positions, such as positions of sensors or cells of a mobile phone network.
- The object positions in the data are arbitrary. This is the case when the positions are received from mobile devices worn by the objects and capable of measuring absolute spatial positions, such as GPS devices.

Therefore, the different cell positions from the data can be used directly as places for the aggregation. To do analysis at a higher spatial scale, we group neighbouring positions to represent the tribunes.

Based on these geometries, the data can be aggregated to *number of visits* to the various sensor locations and *number of flows* among them (see Section 2.2.4.1 for definitions and methods). It should be borne in mind that consecutively visited places in a discontinuous trajectory are not necessarily neighbours in space. As previously described (Section 2.2.4), this first step results in a dual representation of trajectories,

[4]info to the match at http://www.foot-national.com/match-foot-nimesvannes-32912.html, last accessed 08/05/2012

as *sequences of visits* and as *sequences of moves*. The data can be further aggregated in two complementary ways.

First, for each place p_i and time interval Δt, the visits of this place in this interval are aggregated, i.e., the tuples $< o_k, p_i, t_{start}, t_{end} >$ where $\forall t : t_{start} \leq t \leq t_{end}$ and $t \in \Delta t$. The count of the visits and the count of different visitors (o_k) are computed. Hence, each place is characterized by two time series of aggregate values: counts of visits and counts of visitors.

The second way of aggregation is applied to $links$, i.e., pairs of places $< p_i, p_j >$ such that there is at least one move from p_i to p_j. For each link $< p_i, p_j >$ and time interval Δt, the moves from p_i to p_j in this interval are aggregated, i.e., the tuples $< o_k, p_i, p_j, t_0, t_{fin} >$ where $t_{fin} \in \Delta t$ (which means that only the moves that finished within the interval Δt are included). The count of the moves and the count of different objects that moved (o_k) are computed. Hence, each link is characterized by two or more time series of aggregate values: counts of moves. counts of moving objects.

Visualization of these aggregates on maps are described in previous section on *Spatio-Temporal Aggregation*, Section 4.4.2, and depicted in Figure 2.7. These two ways of aggregation support two classes of analysis tasks:

- Investigation of the *presence* of moving objects in different places and the temporal variation of the presence. The presence is expressed by the counts of visits and visitors in the places.

- Investigation of the *flows* (aggregate movements) of objects among different places and the temporal variation of the flows. The flows are represented by the counts of moves and moving objects for the links.

6.4.3.1 Analysis of Presence

Clustering of spatial situations in different time intervals by similarity reduces the workload of the analyst: instead of exploring each situation separately, it is possible to investigate groups of similar situations and to identify events of special interest.

Besides, an appropriate visual representation of the clustering results can disclose the patterns of the temporal variation: whether similar spatial situations occur adjacently or closely in time or may be separated by large time gaps, whether the changes between successive intervals are smooth or abrupt, whether the variation is periodic, etc. For the clustering, the presence situation in each time interval may be represented by a feature vector consisting of the presence values (i.e., the counts of visits and/or visitors) in all places. The results of the clustering are immediately visualized. The centres of the clusters are projected onto a two-dimensional colour space using Sammon's mapping [Sammon 1969]. Thus, a colour is assigned to every time interval. Figure 6.12 depicts the result of the clustering (using self-organizing maps with $k = 8$). The time-period of the match is very characteristic. The lines in the picture represent the recorded number of visitors per scanner, thus it can be seen that during this interval also most of the visitors have been recorded.

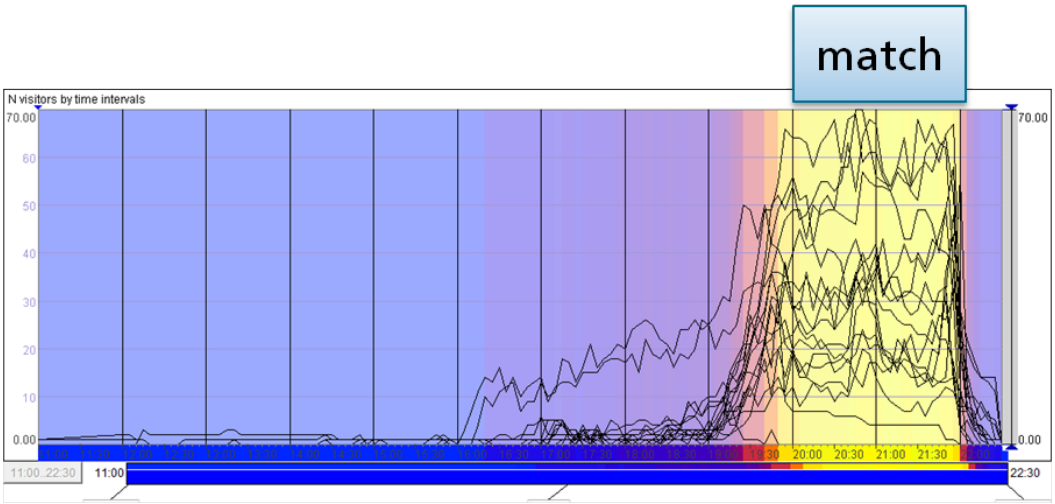

Figure 6.12: Clustering of Presence Situations at Stade des Costières Nîmes (France) [Liebig *et al.* 2013].

6.4.3.2 Analysis of flows

The spatio-temporal variation of the flows is explored analogously to the variation of the presence except that the clustering and visualization tools are applied to the flow situations instead of the presence situations. The flow situation in each time interval is represented by a feature vector consisting of the flow magnitudes (counts of moves and/or moving objects) of the links in this interval.

Figure 6.13 shows results of this clustering after Sammon's projection [Sammon 1969]. Besides the time of the match also the break is recognizable. We utilize the found time intervals for separation of the three phases arrival, departure and match with the temporal boundaries 14:00, 20:00, 21:45 and 22:00.

6.4.3.3 Visual analysis of inner trajectory correlations

Visual exploration of the collected partial trajectories gives indispensable insights of an event. For determination of visitor preferences or identification of potential hazards it is also necessary to discover the dependencies, correlations and patterns among the movements. Therefore, this section tackles the computationally enabled visual exploration of a Bluetooth tracking dataset for inner dependencies which result by the non random movement of the people. Existing approaches e.g. direct database access or usage of a trajectory data warehouse (TDW) [Orlando *et al.* 2007, Raffaetà *et al.* 2011] are unfeasible as the first one requires powerful database hosts and the second pre-aggregates the data which prevents further analysis.

Our approach represents the movement data by an easy to handle descriptive model, namely a Spatial Bayesian Network (SBN) [Liebig *et al.* 2008, Liebig *et al.* 2009]. This probabilistic model denotes the conditional probabilities among visits to discrete locations and thus holds all required information in a compact format for further querying. Afterwards, we utilize the previously trained SBN for visual analysis and

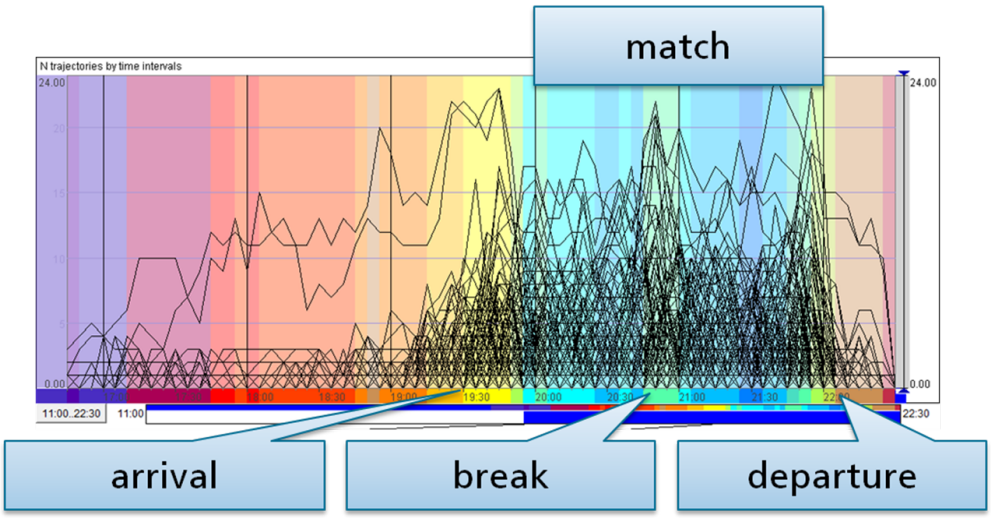

Figure 6.13: Clustering of Flow Situations at Stade des Costières Nîmes (France) [Liebig *et al.* 2013].

depict the probability distributions on three-dimensional thematic maps.

The SBN model is a compact generative representation of the latent correlations within the trajectory dataset (compare Section 4.4.5). The visual user interface, integrated into a Geographic Information System, interacts only with the model and is thus independent of the size of the underlying trajectory database. However, this approach is tailored to one specific analysis task as it extracts patterns early within the analysis process.

First, we construct three-dimensional polygons based on the sensor locations using Dirichlet-Voronoi tessellation. Afterwards, we learn the structure and probability tables of the SBN, which holds a boolean random variable for every polygon. For a single trajectory the random variables associated to the visited places are `true`, and the unvisited ones are `false` (characteristic function).

As described in Section 4.4.5, we applied the improved SSBNL algorithm [Liebig *et al.* 2008, Liebig *et al.* 2009] to the data set using the following parameterization. As the data set is comparably small in its number of variables (usage of 15 sensors implies 15 random variables) for this algorithm, a first pre-sampling step within the trajectories was not necessary. We computed frequent location sets with maximal parity of 4 and a frequency threshold of 5. The Bayesian Network scoring metric we applied was BDeu [Buntine 1991]. In the end we drew 1,000 edge candidates and add negative correlations to the network. The whole Bayesian Network learning took about 1 minute on a standard desktop computer (CPU Intel i7 2GHz, RAM 8GB).

For visualization of the three-dimensional dependencies, we created a Voronoi Dirichlet tessellation of a three-dimensional building model. Both the model and the tessellation geometries were created in Google SketchUp utilizing Ruby scripts for the latter. Materials to the resulting geometries (colour and opacity) are assigned according to the probability distribution computed by the Spatial Bayesian Network. Figure 6.14 depicts the results of the Spatial Bayesian Network for four different queries.

Red colours indicate a high visit probability; blue colours indicate a low probability. The yellow arrows in the picture mark the points of evidence. The Figure 6.14A (in the upper-left corner) depicts the probability distribution given the evidence that the sensor at the ground floor (sensor 34 for comparison in Figure 6.11) has been visited. It is remarkable that the probability on this side of the stadium is high and low in most of the other parts. The places in the other tribunes (at the bottom of the pictures) that possess a relative high probability as well are the VIP rooms and thus visited by the catering staff and prominent visitors from all tribunes after the match ended. In the next step we examine the impact of the staff and prominent guests by change of evidence to a restricted entry within the Spatial Bayesian Network. Results are depicted in Figure 6.14B. All paths that have been used by the catering crew and safety deputies are inked in red which denotes a high probability of movement. The shops possess a relatively high probability. They were located in the uppermost floor of the two towers in the left side of the picture and also in the VIP lounges. As the Bluetooth sensors became subject to vandalism, safety deputies helped us during data collection. Thus it can be seen to the right that they visited sensor location three (top of the upper left tower, compare Figure 6.11) in order to check its presence. In the bottom of Figure 6.14 we combine multiple points of evidence within the query. To the left (picture C) is a visualization of the combined probability of the visitors at the entry to the major tribune and to the VIP entry. The visitors selected by this query distribute among the major tribune and within the VIP rooms. Most likely the untypical movement pattern depicted in picture D was our movement for maintenance of the sensors. The tribune to the left shows a very low probability as it could not be traversed. The tribune on the right was open for traversing before the match began. Thus, our analysis reflects these circumstances and helps to understand movement behavior.

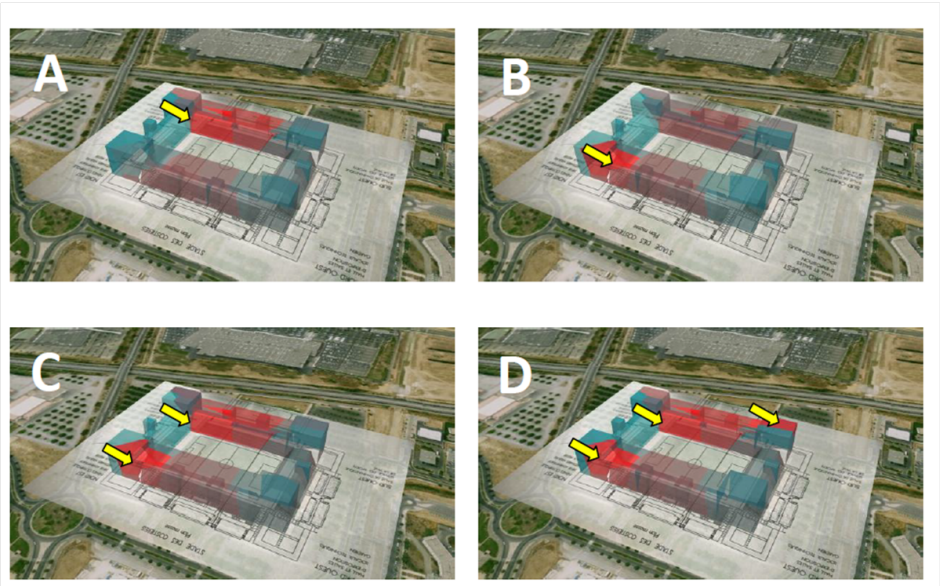

Figure 6.14: Query results - yellow arrows mark location(s) of evidence; blue colour indicates low probability and red indicates high probability of passing by [Liebig *et al.* 2013].

The challenge discussed in this section is the three-dimensionality of the movement data. Thus, we constructed a three-dimensional model of the building where our experiments were conducted. Furthermore, we created three-dimensional geometries of Dirichlet-Voronoi tessellations based on the positions of the sensors. The visualization was integrated in Google Earth using OGC compliant interfaces and a web service. This allows easy integration into other software modules. Another challenge, the dependency analysis of the recorded movement data, was addressed utilizing Spatial Bayesian Networks as an intermediate data structure which holds just the required data instead of complete trajectories. Once the model is built, querying is fast and flexible and overcomes the drawbacks of existing methods that rely on random memory access or aggregation (TDW). The recorded data from the event monitoring application scenario was analyzed in order to identify and reconstruct pedestrian movement. Analysis of the inner-trajectory correlations revealed in-traversable tribunes as well as visitor preferences.

6.4.4 Traffic Quantity Estimation for Event Monitoring

In this section we study our novel pedestrian quantity estimation which incorporates movement patterns (Section 3.7) on the hereby introduced Bluetooth tracking dataset[5].

The quantity estimation method has already been evaluated on synthetic data (Section 3.7.2) as well as real-world data (Section 6.3.6). Thus, it is not the aim to validate the proposed again, but to test whether few sensors could have been omitted for quantity estimation and whether our proposed method is also applicable if the assumptions are not fulfilled completely (details will be given next).

First, the soccer dataset [Liebig & Kemloh Wagoum 2012] is divided into three consecutive time intervals (arrival, match, departure) derived from the clustering of flows in Section 6.4.3.2 (compare Figure 6.13). The time-stamps for splitting are (14:00, 20:00, 21:45, 22:00). During the match there is just low movement of the visitors. Thus, arrival and departure are the most interesting intervals, we therefore conduct our analysis for these intervals.

Hence, the presumption, we made in Definition 6 (Section 3.2) that Kirchhoff's law [Kirchhoff 1845] holds for the traffic network is violated, as in the arrival phase visitors typically enter the building and stay at their seats (vice versa in the departure phase). Furthermore, despite modelling the quantity of moving people (*flow counts*), we are directly going to model the *visit counts* per considered time interval.

Thus, the preprocessing differs from the one in previous applications (synthetic train station data in Section 3.7.2 and tracking data of the visitors to a zoological garden in Section 6.3.6) as follows. The sensor data is not associated to edges, but to vertices. Therefore, the step of line graph [Harary & Norman 1960] mapping (compare Section 3.7) is omitted and values are passed directly to the Gaussian Process Regression. The movement patterns are obtained separately for the considered time

[5] **published with a major contribution of the author in:**
T. Liebig, Z. Xu and M. May. *Incorporating Mobility Patterns in Pedestrian Quantity Estimation and Sensor Placement*. In J. Nin and D. Villatoro, editors, Proceedings of the First International Workshop on Citizen Sensor Networks CitiSens 2012, LNAI 7685, pages 67–80. Springer, 2013

interval, whereas the traffic network is derived from the complete recorded dataset as an aggregation of all sensor transitions (*links*).

Consecutively, the Gaussian process based sensor placement algorithm (Section 3.7.1) is applied to the two datasets: arrival and departure and their movement pattern. The performance of the kernel based sensor placement algorithm is compared to random placement (run 35 times each) by quantity estimation error measured in mean absolute error MAE (Figure 6.15). The Figure depicts the performance for different numbers of sensors, starting from 17 in the left in every step to the right one sensor is omitted.

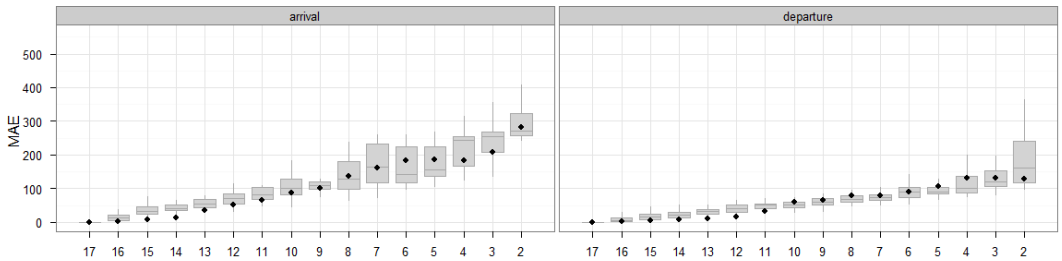

Figure 6.15: MAE for random (grey boxplots) and movement pattern kernel based sensor placement (black dots) [Liebig *et al.* 2013].

In result, the tests reveal, that the GPR method is also applicable if the presumptions are not completely fulfilled. This is justified by the low error the gray boxes (random placement) have, when omitting 'any' sensor. Furthermore, our kernel based sensor placement can outperform random placement: when omitting up to 6 of the applied sensors in the sensor mesh (these are 35%) our placement still outperforms random placement and has an acceptable absolute prediction error of 80 persons (2% of the total number of 3,898 visitors.

6.5 Summary

In this chapter we discussed three real-world applications of the pedestrian quantity estimation methods. The three different discussed scenarios are:

- Billboard Location Evaluation in train stations,
- Visitor Monitoring at the zoo,
- Event Monitoring for a soccer match.

For the billboard location evaluation (in train stations) scenario we based our analysis on previous empirical studies which attest the chosen route of pedestrians in train stations are the quickest path. In the two other scenarios Bluetooth tracking was applied in order to acquire movement patterns. The correlations among the locations are analysed using Visual Analysis, Spatial Bayesian Network representations and quantity estimation methods.

In general, heterogeneous sensor technologies have been applied. For each of the applications we firstly executed a *field study phase* for decisions on sensor selection and

placement and for getting to know the application requirements. The next step was *data preprocessing* and *visual analysis* on the collected sensor data. After data purification, we applied, depending on the application scenario, pedestrian quantity estimation modelling either *presence counts* or *flow counts*.

In the billboard location evaluation scenario we applied our LSR method (Section 3.6) for modelling presences of people in Swiss train stations, i.e. indoor pedestrian movement. The second application scenario focused on visitor movement in a zoological garden. Here we applied our GPR method (Section 3.7) for visitor traffic estimation as well as for sensor placement, incorporating pattern knowledge. Moreover we presented here analysis of *Episodic Movement Data* retrieved from Bluetooth tracking technology, applied in the zoo. The third application discussed the event monitoring scenario at the stadium of Nîmes where we analysed temporal similarities of the *presence* and *flows* of the visitors as well as their co-visit probabilities among the locations in the stadium and successfully applied our methods for sensor placement and quantity estimation for a scenario where our presumption from Chapter 1 on *closed environments* (namely, that Kirchhoff's law is fulfilled for considered time interval) does not hold.

The presented application scenarios are only a few of the possible applications for the *Pedestrian Mobility Analysis System (see Chapter 5)* introduced in this thesis. We have shown in industrial applications that the provided regression methods (LSR and GPR) which incorporate movement patterns are able to predict how many people move or are present at a certain location. We as well applied our innovative methods for sensor placement and obtained good results in practice.

Having the episodic movement data type, we introduced and applied data analysis methods typical for this type of data.

Finally, the positive results on the data sets where the preliminary presumptions are completely violated (i.e. Kirchhoff's law does not hold), are promising for outdoor usage and application to vehicular data, as in these cases just minor violations are expected.

Chapter 7
Discussion

"More importantly, our software worked. I don't just mean that it didn't bump, or that it performed according to the written specifications, or that it was efficient in producing reports. It really worked."

—Eliyahu Moshe Goldratt[1]

Contents

Within this thesis we contributed with analysis methods for episodic movement data, two complementary pedestrian quantity estimation regression methods incorporating movement patterns or expert knowledge and a software system for the application of presented methods. In this chapter, we begin by summarizing the previous contents. Afterwards, we highlight the author's contribution and discuss future research directions. In the closing remarks we address the expected impact of this thesis.

7.1 Synopsis

Analysis of pedestrian mobility is a highly interesting task for quality of service evaluation, location ranking, risk analysis and mobility analysis applications. Particularly modelling of pedestrian quantities gives indispensable insights on visitor preferences and motivations. Therefore, the thesis at-hand focuses on pedestrian quantity estimation methods based on *episodic movement data*. We developed a *System for Pedestrian Mobility Analysis*, presented in Chapter 5. We moreover applied the software system in combination with our contributed quantity estimation methods (Chapter 3) and methods for analysis of episodic movement data (Chapter 4) in real-world industrial scenarios (Chapter 6). The real world applications we apply our methods to, cover the following real world scenarios:

[1] Israeli Physicist, 1947–2011, The Goal: A Process of Ongoing Improvement, [Goldratt, E. M. and Cox, J. 1992]

- *Billboard Location Evaluation* for Swiss train stations: pedestrian quantity estimation in 27 major train stations,
- *Visitor Monitoring* at the zoo of Duisburg: sensor placement and visitor route choice monitoring and
- *Event Monitoring* during a soccer event at Stade des Costières, Nîmes (France): path selection and quantity estimation in spatio-temporal dimension.

Each of the applications requires different traffic modelling, i.e. quantity estimation based on either *moves* or *visits*. Besides conducting experiments on pedestrian quantity estimation, we provide comprehensive analyses of the recorded data sets. Thus, required fundamentals on pedestrian mobility and spatio-temporal data handling are introduced in Chapter 2. The posed questions in each of the applications trigger our method choices.

7.1.1 Summary

In detail, we started in Chapter 2 (after a brief introduction in Chapter 1) by discussing the theoretical basis for our research. We shortly pointed out relevant facts on pedestrian mobility. Then we succinctly reviewed the spatio-temporal geography notions with focus on the state-of-the-art geographical database systems as well as on state-of-the art mobility data types, i.e. the *Episodic Movement Data* with author's contribution.

Since we introduce a pedestrian mobility analysis system, we discussed in the second Chapter the two different classes of mobility models (microscopic and macroscopic ones) and we shortly approached the mobility patterns, as our novel methods (later described in the third Chapter) incorporate movement patterns and return higher accuracy results than similar methods. Moreover, we exposed some of the state-of-the-art pedestrian quantity monitoring techniques, since our algorithms requires this kind of observations.

Chapter 3 discussed the two major contributed methods for pedestrian quantity estimation. Hereby we discussed the related work approaches and we compare them to our own methods. After positioning our contribution in the related work spectrum we described and discussed the two regression approaches: *Least Squares Regression (LSR)* and *Gaussian Process Regression (GPR)*. Moreover we validated and presented the results in this chapter as well.

In the fourth Chapter we focused on the movement pattern analyses based on Bluetooth tracking data. In this chapter we described in detail the Bluetooth tracking technology, as well as its representativeness and we detailed both of the microscopic and macroscopic movement analyses using Bluetooth. For the microscopic movement analysis we briefly reflected the modelling using micro-simulation [Liebig & Kemloh Wagoum 2012] and the monitoring based on Bluetooth radio signal strength [Utsch & Liebig 2012]. In case of macroscopic movement analysis we discussed the acquisition of sequence patterns with Bluetooth tracking technology, spatio-temporal aggregation as well as clustering for (1) presence situations and (2) flow situations. Finally, within the macroscopic movement analysis we dealt with modelling correlations using Spatial Bayesian Networks.

The fifth Chapter strengthens the contribution of this thesis by introducing a practical system for pedestrian mobility analysis supporting the implementation of the previously presented theoretical methods. In this chapter we shortly presented the software development cycle of the pedestrian mobility analysis system tailored for the real-world application scenarios. Firstly, we approached the requirements elicitation phase where the system functionalities have been established to fulfill the application requirements. Secondly we described the system architecture with the composing layers, the interfaces and protocols of communication between them and the temporal interaction between them in a sequence diagram. Lastly, we showed the software integration for our system, in terms of robustness analysis, user interface and integration to the European Emergency Support System [ESS 2010].

The sixth Chapter discussed the successful application of our system into three different real-world industrial scenarios. We divided the chapter into three parts, by roughly providing the knowledge discovery workflow phases [Fayyad *et al.* 1996] for each of the applications. The three industrial cases are firstly, the *Billboard location evaluation* where we applied our method for pedestrian quantity estimation to major Swiss train stations. Secondly, the *Visitor monitoring scenario* in the zoological garden of Duisburg confirmed the benefits of our pedestrian monitoring method incorporating movement patterns, as well as of the sensor placement method. The third application dealt with *Event monitoring* in the stadium of Nîmes during a soccer match. We showed here visual analytics on the sensed data as well as traffic quantity estimation for event monitoring.

7.1.2 Contributions

Throughout this thesis we showed that the proposed methods for pedestrian quantity estimation incorporating movement patterns overcome the weaknesses of the other related works. We have validated our approach against the state-of-the-art related work methods and ground truth and obtained higher accuracy (measured with mean absolute error). We applied the person tracking technology based on Bluetooth in mixed indoor outdoor scenarios. Furthermore, we developed visual analytics methods for analyzing *Episodic Movement Data* in terms of spatial correlations and temporal similarities. We incorporated the functionality for quantitative pedestrian flow analysis in a comprehensive software system for real-world applications and successfully applied the methods in industrial application scenarios. We addressed the posed research questions (compare Section 1.1) within this thesis:

- How to track pedestrians in mixed indoor/outdoor environments?
- How can values on pedestrian quantities be estimated from few empirical measurements?
- At which places should a constrained number of quantity sensors be located?
- How can the developed methods be used in practice?

And contributed with:

- Novel methods for pedestrian quantity estimation and automatic sensor placement which incorporate not just traffic network and sparse sensor readings, but allow incorporation of expert knowledge in form of movement patterns or heuristics on movement patterns.
- Methods for analysis of spatial correlations and temporal similarities contained in *Episodic Movement Data* data, which became prominent, e.g. by proliferation of Bluetooth tracking, RFID tracking, location based social networks or billing records.
- Integration of presented methods in a software framework adjusted to the requirements from real-world application scenarios. Application of the software system to real-world scenarios. Deployment of this software system within the European ESS project [ESS 2010].

We disseminated our contributions in the *data mining* and *machine learning* research field as well as in the *operations research* field by publishing our research in major research volumes, see Section 1.6.

7.2 Future Work

Though the thesis is shaped in a round figure, several new research questions emerged from our analysis which will inspire future research.

Foremost, studies which explore the performance of proposed methods, in case our preconditions are violated, are of interest. We limited our study to the analysis of pedestrian movement in closed environments, as it is easy to validate the estimated results against the ground truth without any side effects (as caused by parking lots for vehicular data). Therefore, it is interesting how our methods perform for vehicular data. For the Gaussian Process Regression we assumed to have noise-free sensors, in practice this is a strong assumption which needs some relaxation.

For the presented sensor placement methods, active learning strategies should be provided and moving sensors are interesting future research directions (e.g. moving Bluetooth sensors, compare [Naini *et al.* 2011]).

Additionally, it should be studied, how the traffic network and the sensor mesh could be defined on different spatial granularities (e.g. cell-based tracking in a mostly unknown area). Therefore, also new sensor technologies should be explored. The rise of location based social networks, the emergent availability of mobile phone data for research and the almost omnipresent video surveillance systems [Köln 2011, webcams.travel 2011] provide valuable input for real-time mobility analysis systems.

Furthermore, the presented tracking technology (Bluetooth tracking) may infringe the privacy of the monitored people. We partly addressed this by application of a hash function to the person identifier. However, in general this is not sufficient as shown in [Monreale *et al.* 2010]. A possible solution also heading for stream processing of recorded data entries could provide usage of sophisticated hashes (sketching techniques) as recently provided to RFID data [Qian *et al.* 2011].

In summary, the future research directions popped up in all the steps we accomplished throughout the thesis. Our major future research focus is the establishment of

a real-time, generative pedestrian mobility model which incorporates real-time sensor readings, dynamic traffic networks, spatio-temporal mobility patterns as well as context information, which is highly needed for real-world applications and for the creation of intelligent environments.

7.3 Closing Remarks

The work on this thesis was driven by real-world applications and projects. The methods performed well under posed preconditions. Thus, the hereby presented methods are already deployed and will find future applications and improvements.

The formalization of a novel data type of episodic spatio-temporal mobility observations, *Episodic Movement Data*, provided a unique term for the research community and will join researchers of different areas, e.g., working on location-based social networks as well as mobile phone logfiles. Additionally, we stimulated the dialog between the research disciplines *Data Mining* and *Operations Research* by comprehensive publication of our research questions and the provided approaches. The study of future research directions and related questions already started, by the author's dissemination of the research results.

As the author I am thrilled by the great scientific impact of the innovative contribution to incorporate movement patterns in pedestrian quantity estimation methods. Personally, I am glad to see, that this thesis was game-changing for the proliferation of Bluetooth tracking and consideration of this technology for future mobility analysis.

Appendix A

Interface Protocols of the System for Pedestrian Mobility Analysis

A.1 Requests to the Software System

The scheme for requests sent to the data analysis module is

```
<?xml version="1" encoding="ISO-8859-1" ?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">
<xs:element name="request">
 <xs:complexType>
  <xs:sequence>
   <xs:element name="screen">
    <xs:complexType>
     <xs:sequence>
      <xs:element name="bounds" type="xs:string"/>
      <xs:element name="center" type="xs:string"/>
      <xs:element name="zoom" type="xs:integer"/>
     </xs:sequence>
    </xs:complexType>
   </xs:element>
   <xs:element name="calendar">
    <xs:complexType>
     <xs:sequence>
      <xs:element name="years">
       <xs:complexType>
        <xs:sequence>
         <xs:element name="year" type="xs:integer"
          maxOccurs="unbounded" minOccurs="0"/>
        </xs:sequence>
       </xs:complexType>
      </xs:element>
      <xs:element name="months">
```

```
        <xs:complexType>
         <xs:sequence>
          <xs:element name="month" type="xs:string"
           maxOccurs="unbounded" minOccurs="0"/>
         </xs:sequence>
        </xs:complexType>
       </xs:element>
       <xs:element name="days">
        <xs:complexType>
         <xs:sequence>
          <xs:element name="day" type="xs:integer"
           maxOccurs="unbounded" minOccurs="0"/>
         </xs:sequence>
        </xs:complexType>
       </xs:element>
       <xs:element name="hours">
        <xs:complexType>
         <xs:sequence>
          <xs:element name="hour" type="xs:integer"
           maxOccurs="unbounded" minOccurs="0"/>
         </xs:sequence>
        </xs:complexType>
       </xs:element>
       <xs:element name="weekdays">
        <xs:complexType>
         <xs:sequence>
          <xs:element name="weekday" type="xs:string"
           maxOccurs="unbounded" minOccurs="0"/>
         </xs:sequence>
        </xs:complexType>
       </xs:element>
      </xs:sequence>
     </xs:complexType>
    </xs:element>
   </xs:sequence>
  </xs:complexType>
 </xs:element>
```

For convenience an example request is stated here:

```
<request>
        <screen>
                <bounds>
                 ((49, 3),
                  (51, 10))
                </bounds>
                <center>(50, 7)</center>
                <zoom>8</zoom>
        </screen>
        <calendar>
                <years>
                        <year>2005</year>
                        <year>2006</year>
                </years>
                <months>
                        <month>April</month>
                        <month>May</month>
                        <month>June</month>
                </months>
                <days>
                </days>
                <hours>
                        <hour>5</hour>
                        <hour>11</hour>
                        <hour>12</hour>
                        <hour>20</hour>
                        <hour>21</hour>
                </hours>
                <weekdays>
                        <weekday>Thursday</weekday>
                        <weekday>Saturday</weekday>
                </weekdays>
        </calendar>
</request>
```

A.2 Replies to the User Interface

The scheme for returned KML 2 reply can be found at http://code.google.com/apis/kml /schema/kml21.xsd. Mainly, we use

`GroundOverlays` to highlight locations and `Placemarks` to store geometries. The geometries are defined as children of the `Placemark` tag similar to this example:

```
<MultiGeometry>
        <Polygon>
          <outerBoundaryIs>
            <LinearRing>
              <coordinates>4, 51, 0
              4, 50, 0
              4, 51, 0
              4, 51, 0
              4, 51, 0
              4, 51, 0</coordinates>
            </LinearRing>
          </outerBoundaryIs>
        </Polygon>
      </MultiGeometry>
```

Time series of predicted values are in the `ExtendedData` tag.

```
      <ExtendedData>
        <Data name="WEEKDAY_Sun">
          <displayName>Sun</displayName>
          <value>295</value>
        </Data>
        <Data name="YEAR_2011">
          <displayName>2011</displayName>
          <value>0</value>
        </Data>
        <Data name="MONTH_May">
          <displayName>May</displayName>
          <value>85</value>
        </Data>
        <Data name="MONTH_Jun">
          <displayName>Jun</displayName>
          <value>110</value>
        </Data>
      </ExtendedData>
```

Appendix B

Brief introduction to Box Plots

The *box plot* was firstly introduced by [Tukey 1970] and it gives a compact visual representation of a distribution. The plot depicts main statistical features: range of the data, median (Q2) and lower (Q1) and upper quartile (Q3). The antennas next to the box are called *whiskers* [McGill *et al.* 1978]. Many variations exist for the definition of the whiskers. In the diagrams presented in this thesis, we apply the $1.5 \cdot IQR$ rule [Frigge *et al.* 1989] (with IQR being the interquartile range $IQR := Q3 - Q1$) as follows: Lower whisker ranges to $Q1 - 1.5 \cdot IQR$, higher whisker ranges till $Q3 + 1.5 \cdot IQR$. Figure B.1 depicts an annotated view of this *box and whisker* plot visualization.

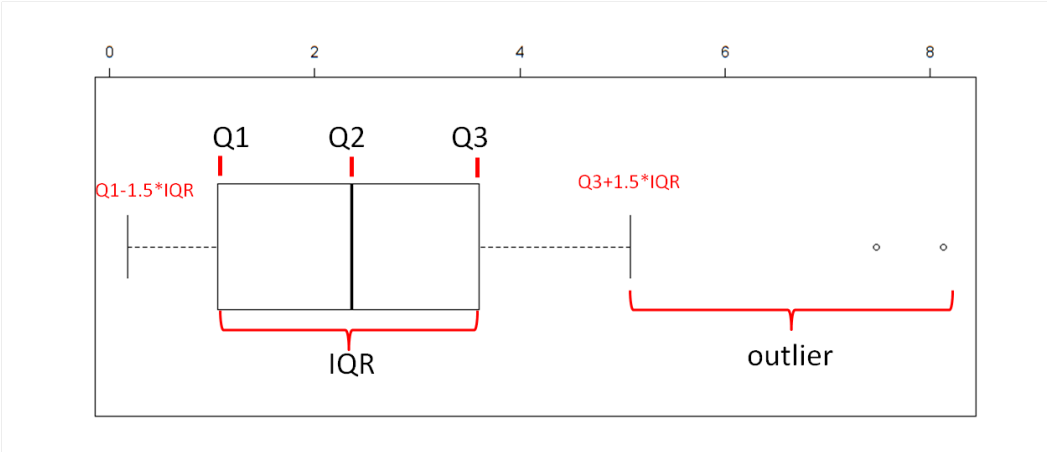

Figure B.1: Example of a box and whisker plot visualization. The box represents lower (Q1), median (Q2) and upper quartile (Q3) of the distribution. Additionally, whiskers and outliers are depicted according to the $1.5 \cdot IQR$ rule.

Appendix C

Tabular Validation Results

The tests of the GPR method were conducted first with synthetic traffic networks (Section 3.7) and afterwards with the dataset collected at the zoological garden in Duisburg (Section 6.3.6).

For every ratio of measured edges (five ratio groups from 10% to 50%) the test is performed 100 times. In case of synthetic data the computation is performed with 100 different traffic networks, in case of the empirical data (just) the sensor positions are swapped. The computations are conducted (using R and python implementations) as follows:

- Firstly, the traffic network is drawn at random from the vertex degree distribution.
- Afterwards, a spatial extent for the traffic network is reconstructed by laying it out.
- A random flow is generated in the traffic network by increasing the path-based frequency between pairwise selected dead-ends (which represent exits of the closed environment).
- Within this network, the sensor locations are defined.
- The (1) frequency values of the sensor locations, (2) the traffic network as well as (3) the previously drawn paths (without their frequency value) are passed to the quantity estimation algorithms.

The process is repeated 100 times for each of the five classes using five algorithms. Therefore for each test 2,500 performance measures were computed and the box-plot visualisation provides a compact view on the results. However, next tables provide the condensed estimation results of the five compared algorithms for the considered scenarios.

Dataset	Algorithm	Min	Q1	Q2	Q3	Max
10%	S-kNN	1541	2964	4150	5128	16238
10%	RL	1541	2752	3212	4893	10343
10%	SE	1541	2964	3947	5128	8664
10%	Diff	1541	2964	3947	5128	8664
10%	Patt	702	2569	2990	5148	11929
20%	S-kNN	1401	2768	3537	4496	11830
20%	RL	801	2329	2887	4185	7943
20%	SE	1100	2321	3035	5244	8862
20%	Diff	1401	3032	4022	4893	12603
20%	Patt	222	1793	2742	4780	10214
30%	S-kNN	1108	2293	2714	3566	5284
30%	RL	536	1833	2400	3380	6571
30%	SE	782	1942	2541	3632	9850
30%	Diff	1185	2840	3584	4435	8776
30%	Patt	61	1310	2190	3145	7734
40%	S-kNN	1185	2084	2509	3012	5292
40%	RL	564	1721	2210	2747	5690
40%	SE	838	1833	2435	3290	7586
40%	Diff	1185	2860	3548	4772	8815
40%	Patt	165	779	1500	2393	6825
50%	S-kNN	862	1566	1925	2197	3473
50%	RL	251	1230	1687	2160	4279
50%	SE	660	1615	2020	2548	5140
50%	Diff	1517	2845	3541	4538	10546
50%	Patt	16	611	1249	1797	4408

Table C.1: Truncated statistical features of the mean absolute error (MAE) distribution depicted in Figure 3.7: Pedestrian quantity estimation on synthetic networks of train stations. Performance is measured by MAE at settings with different ratios of monitored edges (10 to 50 percent). The five methods: GPR with diffusion kernel (Diff), spatial k-nearest neighbour (S-kNN), GPR with trajectory pattern kernel (Patt), GPR with regularized Laplacian (RL) and GPR with squared exponential kernel (SE)

Dataset	Algorithm	Min	Q1	Q2	Q3	Max
10%	S-kNN	1266	1429	1437	1458	1654
10%	RL	1079	1117	1121	1123	1183
10%	SE	1083	1134	1138	1141	1216
10%	CG	1216	1230	1234	1236	1291
10%	Patt	868	966	980	992	1100
10%	**SP**	**467**	**467**	**467**	**467**	**467**
20%	S-kNN	1116	1131	1134	1150	1244
20%	RL	922	973	976	980	991
20%	SE	877	960	964	972	982
20%	CG	1058	1145	1147	1150	1158
20%	Patt	655	741	752	762	790
20%	**SP**	**400**	**400**	**400**	**400**	**400**
30%	S-kNN	883	988	994	1000	1021
30%	RL	813	847	855	859	916
30%	SE	694	793	802	807	822
30%	CG	1046	1050	1052	1056	1166
30%	Patt	475	570	586	596	683
30%	**SP**	**383**	**383**	**383**	**383**	**383**
40%	S-kNN	779	797	810	816	838
40%	RL	702	711	719	721	800
40%	SE	615	625	632	634	700
40%	CG	950	987	989	994	1013
40%	Patt	413	430	431	436	494
40%	**SP**	**383**	**383**	**383**	**383**	**383**
50%	S-kNN	630	666	670	673	683
50%	RL	561	608	611	615	619
50%	SE	458	524	527	528	535
50%	CG	933	957	959	962	983
50%	Patt	325	352	358	369	416
50%	**SP**	**383**	**383**	**383**	**383**	**383**

Table C.2: Truncated statistical features of the mean absolute error (MAE) distribution depicted in Figure 6.9: Quantity estimation performance at the zoo of Duisburg. Performance is measured by MAE at settings with different ratios of monitored edges (10 to 50 percent). The five methods: GPR with diffusion kernel (Diff), spatial k-nearest neighbour (S-kNN), GPR with trajectory pattern kernel (Patt), GPR with regularized Laplacian (RL) and GPR with squared exponential kernel (SE). Values for automatic sensor placement SP (Figure 6.10) are included in red.

Bibliography

[Agrawal & Srikant 1995] R. Agrawal and R. Srikant. *Mining Sequential Patterns*. In Proceedings of the Eleventh International Conference on Data Engineering, pages 3–14, Washington, DC, USA, 1995. IEEE Computer Society. (Cited on pages 36, 37, and 44).

[Alt *et al.* 2009] F. Alt, M. Balz, S. Kristes, A. Sahami S., J. Mennenöh, A. Schmidt, H. Schröder and M. Goedicke. *Adaptive User Profiles in Pervasive Advertising Environments*. In Proceedings of the 3rd European Conference on Ambient Intelligence (AmI'09), volume 5859 of *Lecture Notes in Computer Science*, pages 276–286. Springer, 2009. (Cited on page 3).

[Andrienko & Andrienko 2011] N. V. Andrienko and G. L. Andrienko. *Spatial Generalization and Aggregation of Massive Movement Data*. IEEE Transactions on Visualization and Computer Graphics, vol. 17, no. 2, pages 205–219, 2011. (Cited on pages 23 and 76).

[Andrienko & Andrienko 2012] N. Andrienko and G. Andrienko. *Visual analytics of movement: a review of methods, tools, and procedures*. Journal of Information Visualization, page forthcoming, 2012. (Cited on page 22).

[Andrienko *et al.* 2012] N. Andrienko, G. Andrienko, H. Stange, T. Liebig and D. Hecker. *Visual Analytics for Understanding Spatial Situations from Episodic Movement Data*. KI - Künstliche Intelligenz, pages 241–251, 2012. (Cited on pages xi, xii, xiii, 8, 20, 21, 23, 25, 31, 38, 55, 68, 69, 73, 77, 78, 79, 80, 81, 124, 125, and 130).

[Bartelme 1995] N. Bartelme. Geoinformatik: Modelle, Strukturen, Funktionen. Springer, Berlin, 1995. (Cited on pages xi, 18, and 19).

[Bass *et al.* 2003] L. Bass, P. Clements and R. Kazman. Software Architecture in Practice. SEI Series in Software Engineering. Addison-Wesley, 2003. (Cited on pages 88 and 93).

[Beck *et al.* 2001] K. Beck, M. Beedle, A. van Bennekum, A. Cockburn, W. Cunningham, M. Fowler, J. Grenning, J. Highsmith, A. Hunt, R. Jeffries, J. Kern, B. Marick, R. C. Martin, S. Mellor, K. Schwaber, J. Sutherland and D. Thomas. *Manifesto for Agile Software Development*, 2001. (Cited on page 89).

[Bertozzi *et al.* 2004] M. Bertozzi, A. Broggi, A. Fascioli and A. Tibaldi. *Pedestrian Localization and Tracking System with Kalman Filtering*. In Proceedings of the IEEE Intelligent Vehicles Symposium, pages 584–589, Parma, Italy, 2004. (Cited on pages 3 and 26).

[Blue & Adler 2001] V. J. Blue and J. L. Adler. *Cellular automata microsimulation for modeling bi-directional pedestrian walkways*. Transportation Research Part B: Methodological, vol. 35, no. 3, pages 293–312, March 2001. (Cited on page 33).

[Bluetooth SIG 2004] Bluetooth SIG. *Specification of the Bluetooth System, Volume 0: Core, v2.0*. technical report, November 2004. (Cited on pages 30, 31, 69, and 130).

[Borgers & Timmermans 1986] A. Borgers and H. J. P. Timmermans. *City centre entry points, store location patterns and pedestrian route choice behaviour: A microlevel simulation model*. Socio-Economic Planning Sciences, vol. 20, no. 1, pages 25–31, 1986. (Cited on page 14).

[Bruegge & Dutoit 2010] B. Bruegge and A. H. Dutoit. Object-Oriented Software Engineering: Using Uml, Patterns, and Java. Prentice Hall, 2010. (Cited on pages 88 and 97).

[Bruno & Delmastro 2003] R. Bruno and F. Delmastro. *Design and Analysis of a Bluetooth-based Indoor Localization System*. In In Personal Wireless Communications, IFIP-TC6 8th International Conference, PWC 2003, pages 711–725, 2003. (Cited on pages 3, 30, 67, 68, and 131).

[Buntine 1991] W. Buntine. *Theory Refinement on Bayesian Networks*. In Proceedings of the 7th Annual Conference on Uncertainty in Artificial Intelligence (UAI'91), pages 52–60. Morgan Kaufmann, 1991. (Cited on page 135).

[Chraibi *et al.* 2010] M. Chraibi, A. Seyfried and A. Schadschneider. *The generalized centrifugal force model for pedestrian dynamics*. Physical Review E, vol. 82, page 046111, 2010. (Cited on pages xi, 33, and 34).

[Chraibi *et al.* 2011] M. Chraibi, U. Kemloh, A. Seyfried and A. Schadschneider. *Force-based models of pedestrian dynamics*. Networks and Heterogeneous Media, vol. 6, no. 3, pages 425–442, 2011. (Cited on pages 33, 49, and 72).

[Chu *et al.* 2006] W. Chu, V. Sindhwani, Z. Ghahramani and S. Keerthi. *Relational learning with Gaussian processes*. In Neural Information Processing Systems, 2006. (Cited on page 56).

[Codd *et al.* 1993] E. F. Codd, S. B. Codd and C. T. Salley. *Providing OLAP (On-Line Analytical Processing) to User-Analysis: An IT Mandate*, 1993. (Cited on page 20).

[Costa *et al.* 2007] S. S. Costa, G. Câmara and D. Palomo. *TerraHS: Integration of Functional Programming and Spatial Databases for GIS Application Development*. pages 127–149. 2007. (Cited on page 20).

[De Berg *et al.* 2008] M. De Berg, O. Cheong, M. van Kreveld and M. Overmars. Computational Geometry: Algorithms and Applications. Springer, 3rd edition, April 2008. (Cited on page 35).

[De Raedt 2008] L. De Raedt. Logical and relational learning. Springer, 2008. (Cited on page 56).

[Dedes & Dempster 2005] G. Dedes and A. G. Dempster. *Indoor GPS positioning-challenges and opportunities*. In Proceedings of Vehicular Technology Conference, pages 412–415. IEEE Press, 2005. (Cited on page 29).

[Delone 1934] B. N. Delone. *Sur la sphère vide*. Bulletin of Academy of Sciences of the USSR, no. 6, pages 793–800, 1934. (Cited on pages 42 and 114).

[Demyen & Buro 2006] D. Demyen and M. Buro. *Efficient triangulation-based pathfinding*. In Proceedings of the 21st national conference on Artificial intelligence - Volume 1, AAAI, pages 942–947. AAAI Press, 2006. (Cited on pages 42 and 114).

[Dirichlet 1850] G. L. Dirichlet. *Über die Reduktion der positiven quadratischen Formen mit drei unbestimmten ganzen Zahlen*. Journal für die Reine und Angewandte Mathematik, vol. 40, pages 209–227, 1850. (Cited on pages 23, 42, 76, and 114).

[Doyle 1890] A. C. Doyle. The Sign of Four. Spencer Blackett, 1890. (Cited on page 11).

[Doyle 1969] A. C. Doyle. A Study in Scarlet. John Murray paperback. J. Murray, 1969. (Cited on page 65).

[Ebbutt 2011] G. Ebbutt. Jane's C4I Systems 2011-2012. Jane's C4i Systems. Jane's Information Group, 2011. (Cited on page 129).

[Ellegård *et al.* 1977] K. Ellegård, T. Hägerstrand and B. Lenntorp. *Activity organization and generalization of daily travel: two future alternatives*. Economic Geography, vol. 53, no. 2, pages 167–175, 1977. (Cited on page 37).

[Ellersiek *et al.* 2012] T. Ellersiek, T. Liebig, D. Hecker and C. Körner. *Analyse von raum-zeitlichen Bewegungsmustern auf Basis von Bluetooth-Sensoren*. In Angewandte Geoinformatik 2012 - Beiträge zum 24. AGIT-Symposium Salzburg, pages 260–269, Berlin, 2012. Wichmann. (Cited on pages xiv, 9, 119, 120, 121, 122, 124, and 125).

[Ellersiek 2011] T. Ellersiek. *Bewegungsmuster von Besuchern anhand von Bluetooth Tracking im Duisburger Zoo*. Bachelor thesis, University of Bonn, 2011. (Cited on pages 119, 122, and 124).

[ESS 2010] ESS. *Emergency Support System*. 2010. http://www.ess-project.eu/. (Cited on pages xiii, 6, 91, 101, 103, 104, 106, 107, 129, 143, and 144).

[Fachverband Außenwerbung e.V. 2009] Fachverband Außenwerbung e.V. *Netto-Werbeeinnahmen erfassbarer Werbeträger in Deutschland, 2000-2008 (Net turnover of confirmable advertising media in Gemany, 2000-2008)*, 2009. (Cited on page 113).

[Fayyad *et al.* 1996] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth. *From Data Mining to Knowledge Discovery in Databases*. AI Magazine, vol. 17, pages 37–54, 1996. (Cited on pages 112 and 143).

[FDOT 2012] State of Florida - Departement of Transportation FDOT. *Project Traffic Forecasting Handbook*. technical report, Tallahassee, FL, USA, January 2012. (Cited on pages 46 and 62).

[Florescu *et al.* 2012] S.-C. Florescu, M. Mock, C. Körner and M. May. *Efficient Mobility Pattern Detection on Mobile Devices*. In Proceedings of the ECAI'12 Workshop on Ubiquitous Data Mining, pages 23–27, 2012. (Cited on pages 3 and 66).

[Ford & Fulkerson 1962] L. R. Ford and D. R. Fulkerson. Flows in Networks. Princeton University Press, 1962. (Cited on page 54).

[Friedman *et al.* 1999] N. Friedman, I. Nachman and D. Peér. *Learning Bayesian Network Structure from Massive Datasets: The "Sparse Candidate" Algorithm.* In Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence (UAI'99), pages 206–215. Morgan Kaufmann, 1999. (Cited on page 82).

[Frigge *et al.* 1989] M. Frigge, D. C. Hoaglin and B. Iglewicz. *Some Implementations of the Boxplot*. American Statistician, vol. 43, pages 50–54, 1989. (Cited on page 151).

[Fruchterman & Reingold 1991] T. M. J. Fruchterman and E. M. Reingold. *Graph Drawing by Force-directed Placement*. Software: Practice and Experience, vol. 21, no. 11, pages 1129–1164, 1991. (Cited on page 60).

[Fuentes & Velastin 2001] L. M. Fuentes and S. A. Velastin. *People tracking in surveillance applications*. In Proceedings of the 2nd IEEE International workshop on PETS, 2001. (Cited on page 26).

[Fuller 2009] R. Fuller. Mobile Entity Localization and Tracking in GPS-less Enviroments. In: Tutorial on Location Determination by RF Means. Springer Berlin Heidelberg, 2009. (Cited on pages 28, 29, 30, and 67).

[Galea *et al.* 2004] E. R. Galea, S. Gwynne, P. J. Lawrence, L. Filippidis, D. Blackspields and D. Cooney. *building EXODUS V 4.0 - User Guide and Technical Manual*, 2004. www.fseg.gre.ac.uk. (Cited on page 33).

[Gentili & Mirchandani 2012] M. Gentili and P. B. Mirchandani. *Locating sensors on traffic networks: Models, challenges and research opportunities*. Transportation Research Part C: Emerging Technologies, vol. 24, no. 0, pages 227–255, 2012. (Cited on page 4).

[Getoor & Taskar 2007] L. Getoor and B. Taskar, editors. Introduction to statistical relational learning. The MIT Press, 2007. (Cited on page 56).

[Giannotti & Pedreschi 2008] F. Giannotti and D. Pedreschi. Mobility, Data Mining and Privacy - Geographic Knowledge Discovery. Springer, 2008. (Cited on pages 3, 30, and 66).

[Giannotti *et al.* 2007] F. Giannotti, M. Nanni, F. Pinelli and D. Pedreschi. *Trajectory pattern mining*. In KDD, pages 330–339. ACM, 2007. (Cited on pages 36 and 37).

[Goldenberg & Moore 2004] A. Goldenberg and A. W. Moore. *Tractable Learning of Large Bayes Net Structures from Sparse Data*. In Proceedings of the twenty-first International Conference on Machine learning (ICML'04), pages 44–52. ACM Press, 2004. (Cited on pages 82 and 83).

[Goldratt, E. M. and Cox, J. 1992] Goldratt, E. M. and Cox, J. The goal: a process of ongoing improvement. North River Press, 1992. (Cited on page 141).

[Gong & Wang 2002] X. Gong and F. Wang. *Three Improvements on KNN-NPR for Traffic Flow Forecasting*. In Proceedings of the 5th International Conference on Intelligent Transportation Systems, pages 736–740. IEEE Press, 2002. (Cited on pages 4, 45, 46, 56, and 126).

[Guo & Huang 2011] R. Guo and H. Huang. *Route choice in pedestrian evacuation: formulated using a potential field*. Journal of Statistical Mechanics: Theory and Experiment, vol. 2011, no. 04, page P04018, 2011. (Cited on pages 14 and 34).

[Güting & Schneider 2005] R. H. Güting and M. Schneider. Moving Objects Databases. Morgan Kaufmann, 2005. (Cited on page 19).

[Hagemann & Weinzerl 2008] W. Hagemann and J. Weinzerl. Automatische Erfassung von Umsteigern per Bluetooth-Technologie. In: Nahverkerspraxis. Springer Berlin Heidelberg, 2008. (Cited on pages 31, 68, and 130).

[Hägerstrand 1970] T. Hägerstrand. *What about people in Regional Science?* Papers in Regional Science, vol. 24, no. 1, pages 6–21, 1970. (Cited on pages 17, 37, and 124).

[Hägerstrand 1974] T. Hägerstrand. *Tiidsgeografisk Beskrivning. Syfte och Postulat*. Svensk Geografisk Årsbok, vol. 50, pages 86–94, 1974. (Cited on page 44).

[Hai *et al.* 2012] P. N. Hai, P. Poncelet and M. Teisseire. *GeT_Move: An Efficient and Unifying Spatio-temporal Pattern Mining Algorithm for Moving Objects*. In Proceedings of the 11th International Symposium on Advances in Intelligent Data Analysis, IDA, volume 7619 of *Lecture Notes in Computer Science*, pages 276–288. Springer, 2012. (Cited on page 21).

[Hall *et al.* 2009] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten. *The WEKA data mining software: an update*. SIGKDD Explor. Newsl., vol. 11, no. 1, pages 10–18, November 2009. (Cited on page 78).

[Hallberg *et al.* 2003] J. Hallberg, M. Nilsson and K. Synnes. In Telecommunications, 2003. ICT 2003. 10th International Conference on, title=Positioning with Bluetooth, volume 2, pages 954–958, feb-march 2003. (Cited on pages 31 and 68).

[Harary & Norman 1960] F. Harary and R. Norman. *Some properties of line digraphs*. Rendiconti del Circolo Matematico di Palermo, vol. 9, no. 2, pages 161–168, May 1960. (Cited on pages 56 and 137).

[Hartmann 2010] D. Hartmann. *Adaptive pedestrian dynamics based on geodesics*. New Journal of Physics, vol. 12, page 043032, 2010. (Cited on page 34).

[Hase 2011] H. Hase. Kultur-Markt und Besucherforschung. GRIN Verlag, 2011. (Cited on page 120).

[Helbing & Molnár 1995] D. Helbing and P. Molnár. *Social force model for pedestrian dynamics*. Phys. Rev. E, vol. 51, pages 4282–4286, 1995. (Cited on page 33).

[Helbing 1997] D. Helbing. Verkehrsdynamik: Neue physikalische Modellierungskonzepte. Springer, 1997. (Cited on pages 13, 14, 15, 31, 32, 34, and 46).

[Heliövaara *et al.* 2011] S. Heliövaara, J. Kuusinen, T. Rinne, T. Korhonen and H. Ehtamo. *Pedestrian behavior and exit selection in evacuation of a corridor - An experimental study*. Safety Science, 2011. (Cited on pages 14 and 34).

[Henderson 1971] L. F. Henderson. *The Statistics of Crowd Fluids*. Nature, vol. 229, no. 5284, pages 381–383, february 1971. (Cited on page 32).

[Henderson 1972] L. F. Henderson. *Sexual differences in human crowd motion*. Nature, vol. 240, pages 353–355, 1972. (Cited on page 32).

[Höcker *et al.* 2010] M. Höcker, V. Berkhahn, A. Kneidl, A. Borrmann and W. Klein. *Graph-Based Approaches for Simulating Pedestrian Dynamics in Building Models*. In 8th European Conference on Product & Process Modelling (ECPPM), University College Cork, Cork, Ireland, 2010. http://zuse.ucc.ie/ECPPM/. (Cited on page 35).

[Hoh *et al.* 2012] B. Hoh, T. Iwuchukwu, Q. Jacobson, D. B. Work, A. M. Bayen, R. Herring, J. C. Herrera, M. Gruteser, M. Annavaram and J. Ban. *Enhancing Privacy and Accuracy in Probe Vehicle-Based Traffic Monitoring via Virtual Trip Lines*. IEEE Trans. Mob. Comput., vol. 11, no. 5, pages 849–864, 2012. (Cited on pages 3 and 66).

[Holl & Seyfried 2009] S. Holl and A. Seyfried. *Hermes - an Evacuation Assistant for Mass Events*. inSiDe, vol. 7, no. 1, pages 60–61, 2009. (Cited on page 35).

[Hoogendoorn *et al.* 2002] S. P. Hoogendoorn, P. H. L. Bovy and W. Daamen. *Microscopic Pedestrian Wayfinding and Dynamics Modelling*. In M. Schreckenberg and S. D. Sharma, editors, Pedestrian and Evacuation Dynamics, pages 123–155, 2002. (Cited on pages xi, 12, 13, 14, 15, and 33).

[IEEE 2002] Institute of Electrical and Electronics Engineers and IEEE Computer Society. LAN/MAN Standards Committee and IEEE Standards Board and IEEE Standards Association IEEE. IEEE Standard for Local and Metropolitan Area Networks: Overview and Architecture. IEEE 802 : 2001. IEEE, 2002. (Cited on pages 31, 69, and 131).

[Jacobson & Ng 2004] I. Jacobson and P. Ng. Aspect-Oriented Software Development with Use Cases. Addison-Wesley Professional, 2004. (Cited on page 88).

[Jarvis 1999] R. Jarvis. Romantic Writing and Pedestrian Travel. Palgrave Macmillan, 1999. (Cited on page 1).

[Keim *et al.* 2008] D. Keim, G. Andrienko, J. Fekete, C. Görg, J. Kohlhammer and G. Melançon. *Visual Analytics: Definition, Process, and Challenges Information Visualization*. In Andreas Kerren, John Stasko, Jean-Daniel Fekete and Chris North, editors, Information Visualization, volume 4950 of *Lecture Notes in Computer Science*, chapitre 7, pages 154–175. Springer Berlin / Heidelberg, 2008. (Cited on pages 22 and 74).

[Kemloh Wagoum & Seyfried 2011] A. U. Kemloh Wagoum and A. Seyfried. *Modelling dynamic route choice of pedestrians to assess the criticality of building evacuation*. 2011. arXiv:1103.4080. (Cited on pages 14, 34, and 35).

[Kirchhoff 1845] G. R. Kirchhoff. *Ueber den Durchgang eines elekrischen Stromes durch eine Ebene, insbesondere durch eine kreisförmige*. Annalen der Physik und Chemie, vol. 4, no. 3, pages 497–514, 1845. (Cited on pages 42, 54, and 137).

[Kirchner & Schadschneider 2002] A. Kirchner and A. Schadschneider. *Simulation of evacuation processes using a bionics-inspired cellular automaton model for pedestrian dynamics*. Physica A, vol. 312, pages 260–276, 2002. (Cited on page 33).

[Kirik *et al.* 2009] E. S. Kirik, T. B. Yurgelyan and D. V. Krouglov. *The Shortest Time and/or the Shortest Path Strategies in a CA FF Pedestrian Dynamics Model*. Journal of Siberian Federal University. Mathematics & Physics, vol. 2, no. 3, pages 271–278, 2009. (Cited on page 34).

[Kisilevich *et al.* 2010] S. Kisilevich, D. Keim and L. Rokach. *A Novel Approach to Mining Travel Sequences Using Collections of Geotagged Photos*. In Marco Painho, Maribel Yasmina Santos, Hardy Pundt, William Cartwright, Georg Gartner, Liqiu Meng and Michael P. Peterson, editors, Geospatial Thinking, Lecture Notes in Geoinformation and Cartography, pages 163–182. Springer Berlin Heidelberg, 2010. (Cited on pages 75 and 85).

[Köln 2011] Stadt Köln. *Verkehrs-Webcams in Köln*. http://www.koeln.de/koeln/diedomstadt/verkehr/verkehrskameras/, 2011. (Cited on page 144).

[Kolodziej & Hjelm 2006] K. W. Kolodziej and J. Hjelm. Local Positioning Systems - LBS Applications and Services. CRC Press, Boca Raton, 2006. (Cited on pages 30 and 67).

[Kondor & Lafferty 2002] R. I. Kondor and J. Lafferty. *Diffusion Kernels on Graphs and Other Discrete Input Spaces*. In ICML '02: Proceedings of the Nineteenth International Conference on Machine Learning, pages 315–322, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc. (Cited on pages 50, 57, and 58).

[Kopp *et al.* 2012] C. Kopp, M. Mock and M. May. *Privacy-preserving Distributed Monitoring of Visit Quantities*. In ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, page forthcoming. ACM, 2012. (Cited on page 66).

[Korhonen *et al.* 2008] T. Korhonen, S. Hostikka, S. Heliövaara and H. Ehtamo. *FDS+Evac: An Agent Based Fire Evacuation Model*. Proceedings of the 4th International Conference on Pedestrian and Evacuation Dynamics, february 2008. (Cited on page 33).

[Körner *et al.* 2010] C. Körner, D. Hecker, M. May and S. Wrobel. *Visit Potential: A Common Vocabulary for the Analysis of Entity-Location Interactions in Mobility Applications*. In Marco Painho, Maribel Yasmina Santos, Hardy Pundt, William Cartwright, Georg Gartner, Liqiu Meng and Michael P. Peterson, editors, Geospatial Thinking, volume 0 of *Lecture Notes in Geoinformation and Cartography*, pages 79–95. Springer Berlin Heidelberg, 2010. (Cited on pages 6, 23, 37, and 90).

[Kraak & Ormeling 2003] M. J. Kraak and F. J. Ormeling. Cartography : visualization of spatial data. Pearson Education Limited, second edition edition, 2003. (Cited on pages 24, 78, and 124).

[Krause *et al.* 2006] A. Krause, A. Gupta, C. Guestrin and J. Kleinberg. *Near-optimal sensor placements: Maximizing information while minimizing communication cost*. In IPSN, pages 2–10, 2006. (Cited on page 59).

[Kräußling *et al.* 2008] A. Kräußling, B. Brüggemann, D. Schulz and A. B. Cremers. *People Tracking using Laser Range Scanners and Vision*. In Proceedings of the Fifth International Conference on Informatics in Control, Automation and Robotics, Robotics and Automation 1, pages 29–36. INSTICC Press, 2008. (Cited on pages 3 and 26).

[Kretz & Schreckenberg 2006a] T. Kretz and M. Schreckenberg. *F.A.S.T. - Floor field- and Agent-based Simulation Tool*. technical report physics/0609097, Sep 2006. (Cited on pages 33, 35, and 36).

[Kretz & Schreckenberg 2006b] T. Kretz and M. Schreckenberg. *The F.A.S.T.-Model*. In Samira El Yacoubi, Bastien Chopard and Stefania Bandini, editors, Cellular Automata, volume 4173 of *Lecture Notes in Computer Science*, pages 712–715. Springer Berlin / Heidelberg, 2006. (Cited on pages 33, 35, and 55).

[Kretz 2009] T. Kretz. *Pedestrian traffic: on the quickest path*. Journal of Statistical Mechanics: Theory and Experiment, vol. P03012, mar 2009. (Cited on pages 14 and 34).

[Kuijpers *et al.* 2011] B. Kuijpers, H. J. Miller and W. Othman. *Kinetic space-time prisms*. In Proceedings of the 19th ACM SIGSPATIAL International Symposium on Advances in Geographic Information Systems, ACM-GIS, GIS, pages 162–170. ACM, 2011. (Cited on page 18).

[Lam *et al.* 2006] W. H. K. Lam, Y. F. Tang and M. Tam. *Comparison of two non-parametric models for daily traffic forecasting in Hong Kong*. Journal of Forecasting, vol. 25, no. 3, pages 173–192, 2006. (Cited on pages 4, 47, and 56).

[Leitinger *et al.* 2010] S. Leitinger, S. Gröchenig, S. Pavelka and M. Wimmer. *Erfassung von Personenströmen mit der Bluetooth-Tracking-Technologie.* In Angewandte Geoinformatik 2010 - Beiträge zum 22. AGIT-Symposium Salzburg, pages 220–225, Berlin, 2010. Wichmann. (Cited on pages 31, 68, 85, and 130).

[Lenntorp 1976] B. Lenntorp. Paths in space-time environments: a time-geographic study of movement possibilities of individuals. Meddelanden från Lunds universitets geografiska institution. Royal University of Lund, Dept. of Geography, 1976. (Cited on page 17).

[Leonardi *et al.* 2009] L. Leonardi, G. Marketos, E. Frentzos, N. Giatrakos, S. Orlando, N. Pelekis, A. Raffaetà, A. Roncato, C. Silvestri and Y. Theodoridis. *T-Warehouse: Visual OLAP Analysis on Trajectory Data*. technical report CS-2009-7, Università Ca' Foscari di Venezia, July 2009. (Cited on page 20).

[Li *et al.* 2008] M. Li, S. Konomi and K. Sezaki. *Understanding and modeling pedestrian mobility of train-station scenarios*. In WINTECH, pages 95–96, 2008. (Cited on pages 14, 44, 51, 52, 53, and 113).

[Liebig & Kemloh Wagoum 2012] T. Liebig and A. U. Kemloh Wagoum. *Modelling Microscopic Pedestrian Mobility Using Bluetooth*. In Proc. of the Fourth International Conference on Agents and Artificial Intelligence - ICAART'12, pages 270–275. SciTePress, 2012. (Cited on pages xi, 9, 30, 31, 34, 35, 66, 68, 72, 130, 137, and 142).

[Liebig & Xu 2012] T. Liebig and Z. Xu. *Pedestrian monitoring system for indoor billboard evaluation*. Journal of Applied Operational Research, vol. 4, pages 28–36, 2012. (Cited on pages 8, 51, 53, 90, and 93).

[Liebig *et al.* 2008] T. Liebig, C. Körner and M. May. *Scalable Sparse Bayesian Network Learning for Spatial Applications*. In ICDM Workshops, pages 420–425. IEEE Computer Society, 2008. (Cited on pages 9, 82, 83, 84, 134, and 135).

[Liebig *et al.* 2009] T. Liebig, C. Körner and M. May. *Fast Visual Trajectory Analysis Using Spatial Bayesian Networks*. In ICDM Workshops, pages 668–673. IEEE Computer Society, 2009. (Cited on pages 9, 82, 83, 134, and 135).

[Liebig *et al.* 2010] T. Liebig, H. Stange, D. Hecker, M. May, C. Körner and U. Hofmann. *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. (Cited on pages 8, 53, 114, and 116).

[Liebig *et al.* 2012a] T. Liebig, G. Andrienko and N. Andrienko. *Methods of Analysis of Episodic Movement Data*. In Mobile Tartu, pages 24–25, 2012. (Cited on pages 8, 21, 22, and 73).

[Liebig *et al.* 2012b] T. Liebig, Z. Xu, M. May and S. Wrobel. *Pedestrian Quantity Estimation with Trajectory Patterns*. In Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases ECML PKDD 2012, Part II, LNCS 7524, pages 629–643. Springer, 2012. (Cited on pages xii, xiii, xiv, 8, 40, 43, 48, 55, 60, 61, 62, 120, 121, 127, and 128).

[Liebig *et al.* 2013] T. Liebig, Z. Xu and M. May. *Incorporating Mobility Patterns in Pedestrian Quantity Estimation and Sensor Placement*. In J. Nin and D. Villatoro, editors, Proceedings of the First International Workshop on Citizen Sensor Networks CitiSens 2012, LNAI 7685, pages 67–80. Springer, 2013. (Cited on pages xiv, 8, 21, 55, 82, 83, 112, 121, 129, 131, 134, 135, 136, 137, and 138).

[Liebig 2007] T. Liebig. *Spatio – Temporal Data Mining with Bayesian Networks*. Diploma thesis, Chemnitz University of Technology, 2007. (Cited on pages xi and 19).

[Liebig 2011] T. Liebig. *Trajectory Regression Model for Indoor Pedestrian Flow Analysis on Billboard Evaluation*. In Proc. of the Third International Conference on Applied Operation Research - ICAOR'11, pages 289–300. Tadbir Operational Research Group Ltd., 2011. (Cited on pages 3, 8, 51, and 117).

[Lo *et al.* 2006] S. M. Lo, H. C. Huang, P. Wang and K. K. Yuen. *A game theory based exit selection model for evacuation*. Fire Safety Journal, vol. 41, no. 5, pages 364–369, 2006. (Cited on page 34).

[Marketos *et al.* 2008] G. Marketos, E. Frentzos, I. Ntoutsi, N. Pelekis, A. Raffaetà and Y. Theodoridis. *Building Real World Trajectory Warehouses*. In Proc. 7th International ACM SIGMOD Workshop on Data Engineering for Wireless and Mobile Access (MobiDE'08), 2008. (Cited on page 20).

[Maros & Khaliq 2002] I. Maros and M. H. Khaliq. *Advances in Design and Implementation of Optimization Software*. European Journal of Operational Research, vol. 140, no. 2, pages 322–337, 2002. (Cited on page 1).

[Martin 2009] R. C. Martin. Clean Code: A handbook of agile software craftsmanship. Prentice Hall, 2009. (Cited on page 87).

[Masoud *et al.* 2001] O. Masoud, N. P. Papanikolopoulos and S. Member. *A Novel Method for Tracking and Counting Pedestrians in Real-Time Using a Single Camera*. IEEE Transactions on Vehicular Technology, vol. 50, pages 1267–1278, 2001. (Cited on pages 3 and 26).

[May *et al.* 2008] M. May, D. Hecker, C. Körner, S. Scheider and D. Schulz. *A Vector-Geometry Based Spatial kNN-Algorithm for Traffic Frequency Predictions*. Data Mining Workshops, International Conference on Data Mining, vol. 0, pages 442–447, 2008. (Cited on pages 4, 45, 47, 50, and 59).

[McGill *et al.* 1978] R. McGill, J. W. Tukey and W. A. Larsen. *Variations of Box Plots*. The American Statistician, vol. 32, no. 1, pages 12–16, 1978. (Cited on page 151).

[Meilă 1999] M. Meilă. *An Accelerated Chow and Liu Algorithm: Fitting Tree Distributions to High-Dimensional Sparse Data*. In Proceedings of the Sixteenth International Conference on Machine Learning (ICML'99), pages 249–257. Morgan Kaufmann Publishers Inc., 1999. (Cited on page 83).

[Mínguez *et al.* 2010] R. Mínguez, S. Sánchez-Cambronero, E. Castillo and P. Jiménez. *Optimal traffic plate scanning location for OD trip matrix and route estimation in road networks*. Transportation Research Part B: Methodological, vol. 44, no. 2, pages 282–298, 2010. (Cited on page 4).

[Minkowski 1908] H. Minkowski. *Die Grundgleichungen für die elektromagnetischen Vorgänge in bewegten Körpern*. Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse, vol. 1908, pages 53–111, 1908. (Cited on page 16).

[Minkowski 1909] H. Minkowski. Raum und Zeit, Vortrag, Gehalten auf der 80. Natur-Forscher-Versammlung zu Köln am 21. September 1908. B.G. Teubner, 1909. (Cited on page 18).

[Molnár 1995] P. Molnár. *Modellierung und Simulation der Dynamik von Fußgängerströmen*. Dissertation, Universität Stuttgart, 1995. (Cited on page 33).

[Monreale *et al.* 2010] A. Monreale, G. L. Andrienko, N. V. Andrienko, F. Giannotti, D. Pedreschi, S. Rinzivillo and S. Wrobel. *Movement Data Anonymity through Generalization*. Transactions on Data Privacy, vol. 3, no. 2, pages 91–121, 2010. (Cited on page 144).

[Moore 1965] G. E. Moore. *Cramming more components onto integrated circuits*. Electronics, vol. 38, no. 8, April 1965. (Cited on page 1).

[Naini *et al.* 2011] F. M. Naini, O. Dousse, P. Thiran and M. Vetterli. *Population size estimation using a few individuals as agents*. In Proceedings of the International Symposium on Information Theory, pages 2499–2503. IEEE, 2011. (Cited on pages 85 and 144).

[National Academy of Engineering 2012] National Academy of Engineering. Global Navigation Satellite Systems: Report of a Joint Workshop of the National Academy of Engineering and the Chinese Academy of Engineering. 2012. (Cited on page 29).

[National Imagery and Mapping Agency 2000] National Imagery and Mapping Agency. *Department of Defense World Geodetic System 1984: its definition and relationships with local geodetic systems*. technical report TR8350.2, National Imagery and Mapping Agency, St. Louis, MO, USA, january 2000. (Cited on pages 18, 95, and 98).

[National Institute of Standards and Technology 2002] National Institute of Standards and Technology. Secure Hash Standard. National Institute of Standards and Technology, Washington, 2002. Federal Information Processing Standard 180-2. (Cited on pages 30, 68, and 131).

[National Institute of Standards and Technology 2008] National Institute of Standards and Technology. International System of Units (SI). National Institute of Standards and Technology, Gaithersburg, march 2008. (Cited on page 16).

[Neumann *et al.* 2009] M. Neumann, K. Kersting, Z. Xu and D. Schulz. *Stacked Gaussian Process Learning*. In Proceeding of the 9th IEEE International Conference on Data Mining (ICDM 2009), pages 387–396. IEEE Computer Society, 2009. (Cited on pages 45, 56, 57, 60, and 126).

[Newton *et al.* 1803] I. Newton, A. Motte, W. Davis, W. Emerson and J. Machin. The mathematical principles of natural philosophy by Sir Isaac Newton translated into English by A. Motte to which are added, Newton's system of the world a short comment on, and defence of, the Principia, by W. Emerson; with The laws of the moon's motion according to gravity, by J. Machin . Printed for H.D. Symond, London, A new ed. carefully rev. and corr. by W. Davis. edition, 1803. (Cited on pages 16, 32, and 36).

[Newton 1736] I. Newton. The method of fluxions. Nourse, 1736. (Cited on page 33).

[Ng 2012] M. Ng. *Synergistic sensor location for link flow inference without path enumeration: A node-based approach*. Transportation Research Part B: Methodological, vol. 46, no. 6, pages 781–788, 2012. (Cited on page 4).

[Noschka-Roos 2003] A. Noschka-Roos. Besucherforschung in Museen: Instrumentarien zur Verbesserung der Ausstellungskommunikation. Public understanding of science. Deutsches Museum, 2003. (Cited on page 120).

[Orlando *et al.* 2007] S. Orlando, R. Orsini, A. Raffaetà, A. Roncato and C. Silvestri. *Trajectory Data Warehouses: Design and Implementation Issues*. Journal of Computing Science and Engineering, vol. 1, no. 2, pages 240–261, 2007. (Cited on pages 20, 82, and 134).

[Parkinson 1957] C. N. Parkinson. Parkinson's law : or, the pursuit of progress. Houghton, Boston, 1957. (Cited on page 1).

[Pasquier *et al.* 2008] M. Pasquier, U. Hofmann, F. H. Mende, M. May, D. Hecker and C. Körner. *Modelling and prospects of the audience measurement for outdoor advertising based on data collection using GPS devices (electronic passive measurement system)*. In Proceedings of the 8th International Conference on Survey Methods in Transport, 2008. (Cited on pages 113 and 119).

[Pels *et al.* 2005] M. Pels, J. Barhorst, M. Michels, R. Hobo and J. Barendse. *Tracking people using Bluetooth. Implications of enabling Bluetooth discoverable mode*. technical report, University of Amsterdam, 2005. (Cited on pages 31 and 68).

[Peters & Ennis 2009] C. Peters and C. Ennis. *Modeling groups of plausible virtual pedestrians*. IEEE Computer Graphics and Applications, vol. 29, no. 4, pages 54–63, july 2009. (Cited on pages 14 and 15).

[Picard & de La Hire 1780] J. Picard and P. de La Hire. Traité du nivellement: avec une relation raisonnée de divers nivellemens, & une exposition abrégée de la mesure de la terre. L. Cellot, 1780. (Cited on page 67).

[Qian *et al.* 2011] C. Qian, H. Ngan, Y. Liu and L. M. Ni. *Cardinality Estimation for Large-Scale RFID Systems*. IEEE Trans. Parallel Distrib. Syst., vol. 22, no. 9, pages 1441–1454, 2011. (Cited on page 144).

[Qiang *et al.* 2012] Y. Qiang, M. Delafontaine, M. Versichele, P. De Maeyer and N. Van de Weghe. *Interactive analysis of time intervals in a two-dimensional space*. Information Visualization, 2012. (Cited on page 85).

[R Development Core Team 2009] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. ISBN 3-900051-07-0. (Cited on pages 39 and 53).

[Raffaetà *et al.* 2011] A. Raffaetà, L. Leonardi, G. Marketos, G. L. Andrienko, N. V. Andrienko, E. Frentzos, N. Giatrakos, S. Orlando, N. Pelekis, A. Roncato and C. Silvestri. *Visual Mobility Analysis using T-Warehouse*. International Journal of Data Warehousing and Mining, vol. 7, no. 1, pages 1–23, 2011. (Cited on pages 82 and 134).

[Raney & Nagel 2006] B. Raney and K. Nagel. *An improved framework for large-scale multi-agent simulations of travel behavior*. Towards better performing European Transportation Systems, pages 305–347, 2006. (Cited on page 33).

[Rasmussen & Williams 2006] C. E. Rasmussen and C. K. I. Williams. Gaussian processes for machine learning. The MIT Press, 2006. (Cited on page 56).

[Richardson & Ruby 2007] L. Richardson and S. Ruby. RESTful Web Services. O'Reilly Series. O'Reilly Media, Incorporated, 2007. (Cited on pages 91, 101, and 129).

[Rigoutsos & Floratos 1998] I. Rigoutsos and A. Floratos. *Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm*. Bioinformatics, vol. 14, no. 1, pages 55–67, 1998. (Cited on page 75).

[Rindsfüser 2005] G. Rindsfüser. *Simulation des Aktivitätenplanungsprozesses mit Methoden der Künstlichen Intelligenz*. Straßenverkehrstechnik, vol. 49, no. 3, pages 145–153, 2005. (Cited on pages 51 and 52).

[Sammon 1969] J. W. Sammon. *A Nonlinear Mapping for Data Structure Analysis*. IEEE Transaction on Computers, vol. 18, no. 5, pages 401–409, may 1969. (Cited on pages 78, 133, and 134).

[Schadschneider *et al.* 2009] A. Schadschneider, W. Klingsch, H. Kluepfel, T. Kretz, C. Rogsch and A. Seyfried. Encyclopedia of Complexity and System Science, volume 5, chapitre Evacuation Dynamics: Empirical Results, Modeling and Applications, pages 3142–3176. Springer, Berlin, Heidelberg, 2009. (Cited on page 33).

[Schmidt *et al.* 2011] M. Schmidt, D. Kim and S. Sra. *Projected Newton-type Methods in Machine Learning*. In S. Sra, S. Nowozin and S. J. Wright, editors, Optimization for Machine Learning, Neural Information Processing Series, chapitre 11, pages 297–321. 2011. (Cited on page 55).

[Schmutzer 1989] E. Schmutzer. Relativitätstheorie - aktuell : ein Beitrag zur Einheit der Physik. Teubner, Leipzig, 4 edition, 1989. (Cited on pages 16 and 17).

[Scholz 2009] A. Scholz. Besucherforschung an Museen. GRIN Verlag, 2009. (Cited on page 120).

[Schulz *et al.* 2003] D. Schulz, W. Burgard, D. Fox and A. B. Cremers. *People Tracking with Mobile Robots Using Sample-based Joint Probabilistic Data Association Filters*. International Journal of Robotic Research, vol. 22, no. 2, pages 99–116, 2003. (Cited on pages 3 and 26).

[Seitner & Hanbury 2006] F. H. Seitner and A. Hanbury. *Fast pedestrian tracking based on spatial features and colour* . In Ondřej Chum and Vojtěch Franc, editors, CVWW'06: Proceedings of the Computer Vision Winter Workshop 2006 , pages 105–110, Prague, Czech Republic, February 2006. Czech Society for Cybernetics and Informatics. (Cited on page 26).

[Selby & Kockelman 2013] B. Selby and K. M. Kockelman. *Spatial prediction of traffic levels in unmeasured locations: applications of universal kriging and geographically weighted regression*. Journal of Transport Geography, vol. 29, pages 24–32, May 2013. (Cited on page 50).

[Seyfried *et al.* 2010] A. Seyfried, M. Chraibi, U. Kemloh, J. Mehlich and A. Schadschneider. *Runtime Optimization of Force Based Models within the Hermes Project*. In Pedestrian and Evacuation Dynamics 2010, pages 363–373. Springer, 2011, 2010. (Cited on page 35).

[Smola & Kondor 2003] A. Smola and R. Kondor. *Kernels and Regularization on Graphs*. In Proc. Conf. on Learning Theory and Kernel Machines, pages 144–158, 2003. (Cited on page 50).

[Spinello *et al.* 2010] L. Spinello, K. O. Arras, R. Triebel and R. Siegwart. *A Layered Approach to People Detection in 3D Range Data.* In AAAI, pages 1625–1630. AAAI Press, 2010. (Cited on page 26).

[Spinello *et al.* 2011] L. Spinello, M. Luber and K. O. Arras. *Tracking People in 3D Using a Bottom-Up Top-Down Detector.* In Proc. of The International Conference in Robotics and Automation (ICRA), pages 1304–1310. IEEE, 2011. (Cited on pages xi, 26, and 27).

[Stange *et al.* 2011] H. Stange, T. Liebig, D. Hecker, G. Andrienko and N. Andrienko. *Analytical Workflow of Monitoring Human Mobility in Big Event Settings using Bluetooth*. In ISA 2011, pages 51–58. ACM, 2011. (Cited on pages 9, 21, 30, 31, 68, 112, 124, and 130).

[Starke 2005] G. Starke. Effektive Software-Architekturen. b-Agile. Hanser, 2005. (Cited on pages 89 and 98).

[Stiftung Werbestatistik Schweiz 2009] Stiftung Werbestatistik Schweiz. *Werbeaufwand Schweiz, Erhebungsjahr 2008 (Advertising expenditure Switzerland, survey year 2008)*, 2009. (Cited on page 113).

[Swiss Poster Research Plus 2010] SPR+ Swiss Poster Research Plus. *Fraunhofer IAIS Project*. http://www.iais.fraunhofer.de/spr.html, 2010. http://www.iais.fraunhofer.de/spr.html. (Cited on pages xii, xiii, 54, 113, 115, 116, and 117).

[Tanyimboh & Templeman 1993] T. T. Tanyimboh and A. B. Templeman. *Calculating maximum entropy flows in networks*. Journal on Operation Research, vol. 44, no. 4, pages 383–396, 1993. (Cited on pages 46 and 52).

[Teichman & Thrun 2012] A. Teichman and S. Thrun. *Tracking-based semi-supervised learning*. International Journal of Robotic Research, vol. 31, no. 7, pages 804–818, 2012. (Cited on pages 3 and 26).

[Thompson 1994] P. A. Thompson. *Developing new techniques for modelling crowd movement*. Phd thesis, University of Edinburgh, 1994. (Cited on page 33).

[Tobler 1970] W. Tobler. *A Computer Movie Simulating Urban Growth in the Detroit Region*. Economic Geography, vol. 46, no. 2, pages 234–240, 1970. (Cited on page 45).

[Torge 2001] W. Torge. Geodesy. W. de Gruyter, 2001. (Cited on page 68).

[Tukey 1970] J. W. Tukey. Exploratory Data Analysis. Numeéro 1 de Exploratory Data Analysis. Addison Wesley Publishing Company, 1970. (Cited on page 151).

[Utsch & Liebig 2012] P. Utsch and T. Liebig. *Monitoring Microscopic Pedestrian Mobility Using Bluetooth*. In Proceedings of the 8th International Conference on Intelligient Environments, pages 173–177. IEEE Press, 2012. (Cited on pages 9, 31, 66, 68, 72, 85, and 142).

[Utsch 2011] P. Utsch. *Modellierung von Fußgängerströmen in Gebäuden unter der Verwendung von Bluetooth-Tracking-Daten*. Student resarch project, Bonn-Rhein-Sieg University of Applied Sciences, 2011. (Cited on page 72).

[Versichele *et al.* 2012a] M. Versichele, R. Huybrechts, T. Neutens and N. Van de Weghe. *Intelligent Event Management with Bluetooth Sensor Networks*. In Proceedings of the 8th International Conference on Intelligient Environments, pages 311–314. IEEE Press, 2012. (Cited on pages 69 and 85).

[Versichele *et al.* 2012b] M. Versichele, T. Neutens, M. Delafontaine and N. Van de Weghe. *The use of Bluetooth for analysing spatiotemporal dynamics of human movement at mass events: A case study of the Ghent Festivities*. Applied Geography, vol. 32, no. 2, pages 208–220, march 2012. (Cited on pages 85 and 130).

[Viger & Latapy 2005] F. Viger and M. Latapy. *Efficient and Simple Generation of Random Simple Connected Graphs with Prescribed Degree Sequence*. In Lusheng Wang, editor, Computing and Combinatorics, volume 3595 of *Lecture Notes in Computer Science*, chapitre 45, pages 440–449. Springer Berlin/Heidelberg, 2005. (Cited on page 60).

[Voronoï 1908] G. Voronoï. *Nouvelles applications des paramètres continus à la théorie des formes quadratiques. Deuxième mémoire. Recherches sur les parallélloèdres primitifs.* Journal für die reine und angewandte Mathematik (Crelle's Journal), no. 134, pages 198–287, December 1908. (Cited on pages 22, 23, 42, 76, and 114).

[Wang & Kockelmann 2009] X. Wang and K. M. Kockelmann. *Forecasting Network Data: Spatial Interpolation of Traffic Counts from Texas Data*. Journal of the Transportation Research Board, vol. 2105, no. 13, pages 100–108, 2009. (Cited on pages 4 and 47).

[webcams.travel 2011] webcams.travel. *WebCam Karte*. http://de.webcams.travel/map/, 2011. (Cited on page 144).

[Weidmann 1993] U. Weidmann. *Transporttechnik der Fussgänger - Transporttechnische Eigenschaften des Fussgängerverkehrs (Literaturstudie)*. Literature Research 90, Institut für Verkehrsplanung, Transporttechnik, Strassen- und Eisenbahnbau IVT an der ETH Zürich, ETH-Hönggerberg, CH-8093 Zürich, March 1993. in German. (Cited on page 12).

[Wikimedia Commons] Wikimedia Commons. *Example of a light cone, by Stib at en.wikipedia (Transferred from en.wikipedia.)*

[GFDL (www.gnu.org/copyleft/fdl.html) or CC-BY-SA-3.0 (http://creativecommons.org/licenses/by-sa/3.0/)]. (Cited on pages xi and 16).

[Witten & Frank 2005] I. H. Witten and E. Frank. Data mining : practical machine learning tools and techniques. Morgan Kaufman, Amsterdam, 2 edition, 2005. (Cited on pages 45 and 46).

[Woodings *et al.* 2001] R. Woodings, D. Joos, T. Clifton and C. D. Knutson. *Rapid Heterogeneous Connection Establishment: Accelerating Bluetooth Inquiry Using IrDA*. 2001. (Cited on pages 30, 68, and 131).

[Yang & Zhou 1998] H. Yang and J. Zhou. *Optimal traffic counting locations for origin–destination matrix estimation*. Transportation Research Part B: Methodological, vol. 32, no. 2, pages 109–126, 1998. (Cited on page 4).

[Yassin-Kassab *et al.* 1999] A. Yassin-Kassab, A. B. Templeman and T. T. Tanyimboh. *CALCULATING MAXIMUM ENTROPY FLOWS IN MULTI-SOURCE, MULTI-DEMAND NETWORKS*. Journal on Engineering and Optimization, vol. 31, no. 6, pages 695–729, 1999. (Cited on pages 46 and 52).

[Yu *et al.* 2005] W. J. Yu, L. Y. Chen R. Dong and S. Q. Dai. *Centrifugal force model for pedestrian dynamics*. Phys. Rev. E, vol. 72, no. 2, page 026112, august 2005. (Cited on page 33).

[Yu *et al.* 2006] K. Yu, W. Chu, S. Yu, V. Tresp and Z. Xu. *Stochastic relational models for discriminative link prediction*. In Neural Information Processing Systems, 2006. (Cited on page 56).

[Zhao & Park 2004] F. Zhao and N. Park. *Using Geographically Weighted Regression Models to Estimate Annual Average Daily Traffic*. Journal of the Transportation Research Board, vol. 1879, no. 12, pages 99–107, 2004. (Cited on pages 4, 45, 46, and 126).

[Zhu *et al.* 1997] C. Zhu, R. H. Byrd, P. Lu and J. Nocedal. *Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization*. ACM Trans. Math. Softw., vol. 23, no. 4, pages 550–560, December 1997. (Cited on page 53).