3D Indoor Movement Analysis and Visualization Utilizing Bluetooth Tracking and Spatial Bayesian Networks

Thomas Liebig and Gennady Andrienko and Natalia Andrienko

Fraunhofer IAIS; Schloss Birlinghoven; 53754 Sankt Augustin
{thomas.liebig, gennady.andrienko, natalia.andrienko}@iais.fraunhofer.de

Abstract

Visual analysis of trajectory data became a common approach during the past years. Considering advances in pedestrian tracking technology, Bluetooth tracking data received recent attention. In this paper we present a fast, model-based approach for computationally enabled visual exploration of location dependencies in Bluetooth tracking data sets. Existing approaches are not suitable for visual dependency analysis as the size and complexity of trajectory data constrain ad-hoc and advance computations. Also recent developments in the area of trajectory data warehouses cannot be applied because the spatial correlations are lost during trajectory aggregation. Our approach builds a compact Spatial Bayesian Network model, which represents the dependency structures of the data. The user queries are answered using this intermediate model instead of the complete data set. Visualization is connected by Open Geographic Consortium compliant protocols and uses 3D Dirichlet-Voronoi tessellation. This paper presents the approach and applies it on a soccer match dataset.

1 Introduction

Major airports, arenas and stadiums are designed to attract thousands or billions of visitors each year. Whereas one trend is to build larger infrastructures (airports, stadiums) another trend is the growing visitor number at events. In the last years, this hazardous development led to devastating disasters (e.g. Loveparade 2010). Thus, visitor monitoring in complex facilities became an important subject. But understanding the movement behavior, identification of attractors and distractors, determination of waiting times, as well as localization of congestions and bottle-necks gives also insights on visitor preferences and motivations at a particular site or event. Knowing such detailed information on indoor pedestrian behavior 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 [6] became the state-of-the-art method for combined indoor outdoor monitoring of pedestrian movement [4, 10, 12, 14, 22, 23].

Visual exploration of the collected partial trajectories gives indispensable insights of an event [9, 14]. 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 work tackles the computationally enabled visual exploration of a massive 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) [13, 20] are unfeasible as the first one requires powerful database hosts and the second pre-aggregates the data which prevents further analysis. Our proposed method contains two stages. We represent the massive movement data by an easy to handle descriptive model, namely a Spatial Bayesian Network (SBN) [16]. 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. In step 2 we utilize the previously trained SBN for visual analysis and depict the probability distributions on three-dimensional thematic maps. The latter is integrated in Google Earth using web services and Open Geographic Consortium (OGC) compliant data formats.

The remainder of the paper proceeds as follows. The upcoming section 2 gives an overview on related Bluetooth tracking work and introduces trajectory dependency models. We give a brief summary of Spatial Bayesian Networks in section 3. Our approach is applied in section 4 where we conduct experiments on a soccer match data set. We conclude in the final section 5.

2 Related Work

Analysis of movement data recorded by established tracking technologies (e.g. Global Positioning System GPS) has become research focus during the last years. Visual analysis became a natural approach due to the spatio-temporal nature of the movement data. Also Bluetooth tracking data, as one characteristic type of spatio-temporal movement records, has received notable attention from the area of visual analytics recently [22, 4]. In this paper we consider the visualization of dependencies within cell based (Bluetooth tracking) data. A number of different research areas have contributed to the analysis of geographic data. Next to geostatistics and geographic information systems and science, database technology and data mining play a major role in the development of analysis methods for large spatial and spatio-temporal data sets. While spatial database technology and spatial data mining have become well-established parts in their respective research areas, a number of gaps in covering the task space by appropriate computational methods have been identified [3].

Recently, two approaches for aggregation and analysis of large sets of trajectory data have been published. Both approaches rely on a database-side aggregation of the data prior to data analysis and use the Visual Analytics Toolkit (VAT) for visualization. The Visual Analytics Toolkit [2] is a software system for interactive visual analysis of spatially and temporally referenced data in context. The first approach by Andrienko and Andrienko [1] relies on the user to perform aggregation using standard database functions. VAT provides a direct database access, and the user can load tables with previously aggregated data. Only if the data set is small enough to fit into random access memory, so-called dynamic aggregators can be applied directly within the VAT. The second approach by Leonardi et al. [13] performs aggregation using a Trajectory Data Warehouse TDW. TDW [17, 20] have recently been developed and are first steps into OLAP (Online Analytical Processing) analysis of trajectory data. The TDW stores aggregated data at a given level of resolution. VAT interacts with the TDW to allow for visually aided OLAP, e.g. roll-up and drill-down operations for graphically selected areas.

However, these two approaches are not suitable for complex visual dependency analysis. First, the exponential number of location subsets prohibits advance computation and ad-hoc calculation of dependencies may take too long for large data sets. Second, TDW naturally do not keep the identity of trajectories during aggregation, which makes inference of location dependencies impossible. In contrast to the above approaches, the process in [15] is not based on aggregated data but on a generative probabilistic model of the data. The model is a compact 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. However, this approach is tailored to one specific analysis task as it extracts patterns early within the analysis process. In contrast, approaches 1 and 2 are flexible with respect to possible analysis questions, because the selection and control of analyses resides with the user in the upper most level. We inspire our method by [15]. Thus, we scale it up to a third dimension, construct three-dimensional polygons and connect all the software components by Open Geogrphic Consortium (OGC) compliant protocols. The next section describes Spatial Bayesian Networks which is a dependency model among spatial objects and is utilized by our approach for compact representation of inner trajectory correlations. Thus, it builds the core of our approach.

3 Spatial Bayesian Networks

Visual Analysis of visitor behavior is a natural and promising approach to understand visitor preferences and reveal patterns among the movements. Existing querying techniques have drawbacks in handling the massive three-dimensional movement recordings (see section 2). 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.

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. These co-occurrences can be expressed as conditional probability to visit an arbitrary location given that another (set of) location(s) is visited within the same movement as well. More formally, given a finite universal set L of discrete geographic locations, a set $L^+ \subseteq L$ containing locations that are visited with certainty and a set $L^- \subseteq L \setminus L^+$ containing locations that are not visited with certainty within a trajectory, we can specify the location dependency of an arbitrary location $l \in L$ by the probability $P(l \mid L^+; \neg L^-)$. The sets L^+ and L^- are also called positive and negative evidence, respectively. Bayesian Networks are a common approach to model such dependencies. It combines random variables $\boldsymbol{X} = X_1, X_2 \ldots X_n$ by a directed acyclic graph which denotes dependencies as well as conditional independencies. Thus, we assign to every location l in L exactly one boolean random variable X_i of $\boldsymbol{X}$ which is TRUE, iff a particular trajectory passes by and FALSE otherwise. At each vertex (i.e. each random variable X_i) a common probability table denotes the dependency of this random variable X_i from its parents (which are the random variables that are connected to X_i by directed edges) $P(X_i{=}x_i|parents(X_i))$. The joint probability distribution for all random variables $\mathbf{X}$ is then given by the following equation:

$$p(\boldsymbol{X} = \boldsymbol{x}) = p(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \prod_{i=1}^{n} p\,(X_i = x_i | parents(X_i)).$$

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) [16] as this was especially designed to meet the demands of spatial data mining. SSBNL combines the advantages of the Sparse Candidate Learning [7] and the Screen Based Network Search [8]. It bounds the number of possible ancestors in the network by pre-sampling a given sparseness in the database, and bounds the edge set to most significant dependencies by processing only frequent item sets similar to [8]. This is done in a two-step algorithm: First, the algorithm pre-samples within each route a set of maximal k distinct locations. The sampling is uniformly distributed among the trajectory; i.e. every discrete location contained in the trajectory is drawn with the same probability. Afterwards, frequent variable sets become sampled 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 re-compute 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 [15], 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 needs also 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 visitor passes location A it is unlikely to pass location B within the same movement". 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 1).

Require: 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
Ensure: 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

Algorithm 1: Scalable Sparse Bayesian Network Learning Algorithm (SSBNL)

This pair wise approach is reasonable as shown in [18]. The complete network learning Algorithm is summarized in algorithm 1. After the dependencies among the locations are extracted utilizing Spatial Bayesian Networks, a graphical user interface is required to inspect the correlations of different locations within the underlying trajectory set. Our approach is to embed this as three-dimensional layer into Google Earth by Open Geographic Consortium (OGC) compliant Keyhole Markup Language (KML) interfaces. Our extension for Google Earth consists of two parts: a web service for KML creation and update, and a mediator script to trigger Bayesian network calculations. Each user query needs to execute Bayesian inference on the Spatial Bayesian Network according to the given evidence. In order to keep this part independent of the currently used GIS, we create a separate mediator script written in the language R [22]. Thus, we may easily use other geographic information systems or access the learned Spatial Bayesian Network from different applications written in R script as well. An advantage of the scripting language R is its large collection of statistical analysis packages and references. In our case we use the Bayesian Network data structure defined in the deal package [5]. The data exchange between the user interface and mediator script is implemented using files. The control flow and synchronization of the execution is solved calling a shell execution command at each computation request sent by the web service. This means, any computation cycle starts a single R process that reads the evidence from a file and stores the inference results in a different file. The web service uses this file and creates its output accordingly. In order to prevent long import times of the Spatial Bayesian Network every time a new R process is created, we store the complete R workspace with all objects (including the Bayesian Network) as the default workspace. The workspace is read very fast at startup and written after execution automatically. Combining all parts, our fast query tool based on Bayesian Networks consists of a layered structure. The architecture offers several possibilities for the independent exchange of components, which is important for future development and reusability. The spatial dependency model may be accessed by other tools and the Spatial Bayesian Network may also be replaced by a more accurate one or even a complete different dependency model.

The method for analysis of inner-trajectory dependencies, presented in the previous section, enables us to study relations among multiple locations in movement data sets performed in the next section.

4 Experiments

In this section, we conduct tests of our approach on a dataset collected through privacy preserving Bluetooth tracking technology [14]. For data collection a mesh of 15 sensors has been deployed among a soccer stadium (Stade des Cosières, Nîmes at France) during a soccer[1] match on 05.08.2011. The collected dataset will be published after acceptance of this work. The three-dimensional sensor placement is depicted in Figure 1. All Bluetooth enabled devices (e.g. smartphones or intercoms) passing one of the sensors (more precisely its footprint) trigger the creation of a datalog entry consisting of the timestamp, the sensor identifier (which denotes the position), the radio signal strength and a hashed identifier for this particular device [19]. Whenever a device passes multiple sensors, it becomes re-detected and transition times as well as movement patterns can be reconstructed. We recorded 47,589 data points from 553 different devices at 15 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 3898[1] (this official visitor number does not contain the people which worked there), have been recorded during the period of the match, thus we expect the dataset to be representative.

Our analysis is inspired by the two step workflow presented in [22]. Whereas the *field study* phase was performed during (1) survey design and (2) data collection we conduct the second *knowledge discovery* phase within the (3) data preparation (4) data mining. In contrast to the existing workflow,

[1] http://www.foot-national.com/match-foot-nimesvannes-32912.html, last accessed 29.02.2012

the last step, visualization, becomes a loop, where visualization stimulates user interaction which itself triggers the re-computation of the data mining model.

We applied the improved SSBNL algorithm (see Algorithm 1) 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 [11]. In the end we drew 1000 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).

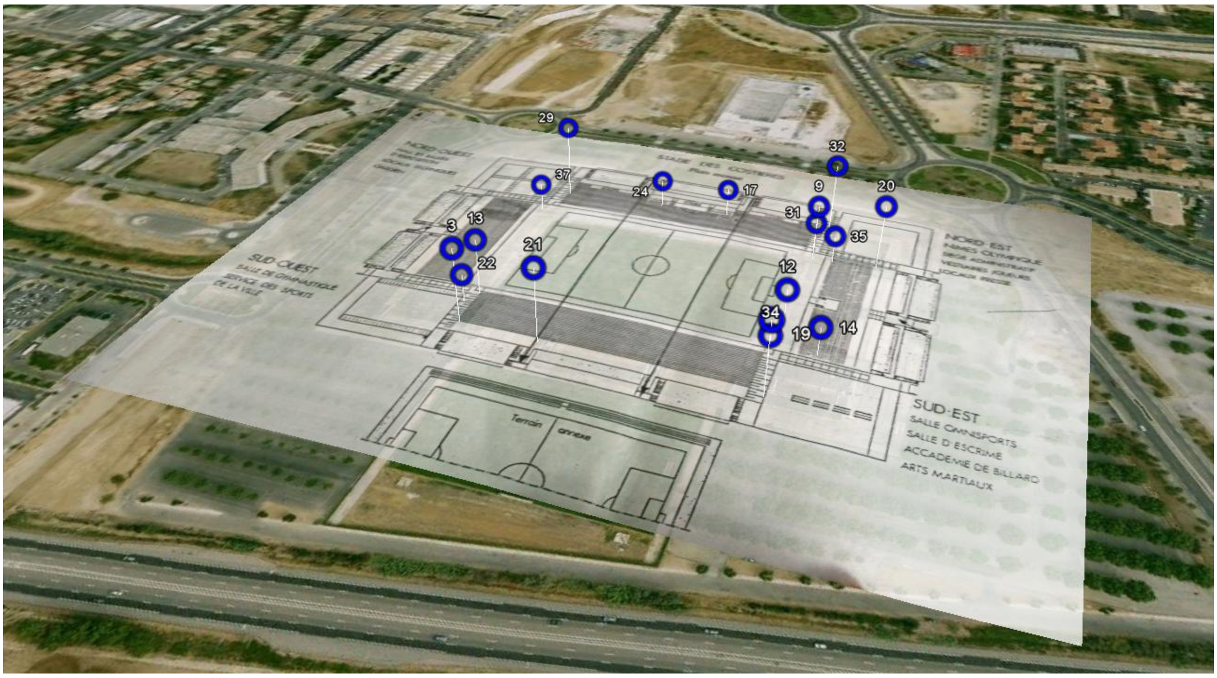

Figure 1: 3D Sensorplacement at Stade des Costières, Nîmes (France) blue circles mark the sensor positions, the number denote their senor identifiers

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 (color and opacity) are assigned according to the probability distribution computed by the Spatial Bayesian Network. Figure 2 depicts the results of the Spatial Bayesian Network for four different queries. Red colors indicate a high visit probability; blue colors indicate a low probability. The yellow arrows in the picture mark the points of evidence. The picture A (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 1) 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 picture B. 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 1) in order to check its presence. In the bottom of figure 2 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. By further of addition of sensor location three the places considered so far reach their highest conditional probability. Most likely this 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.

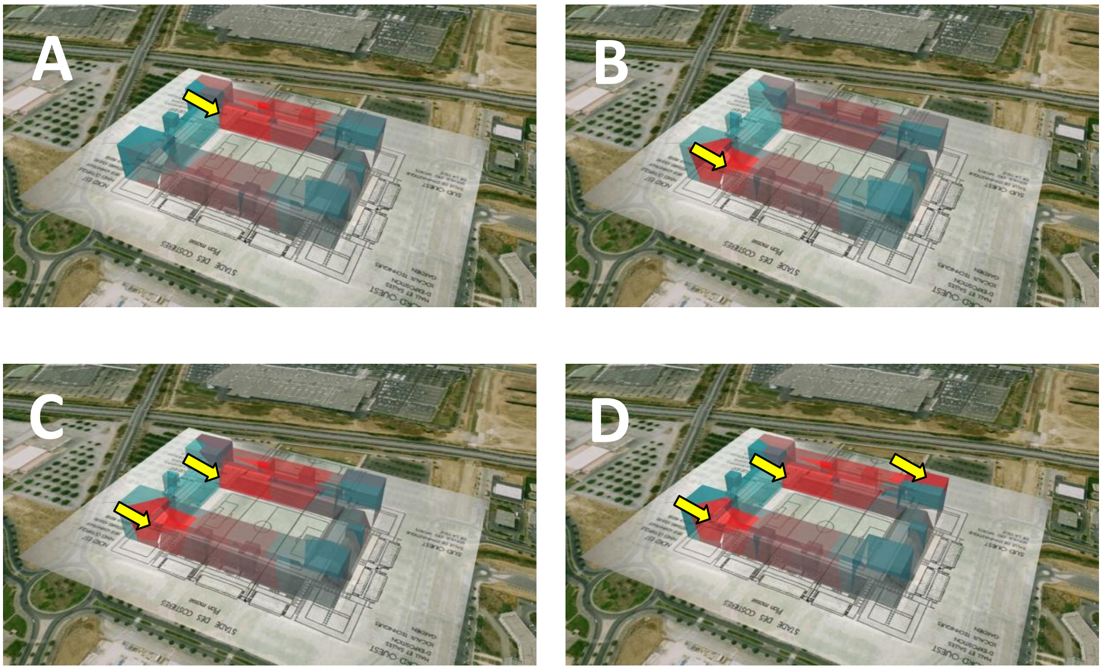

Figure 2: Query results - yellow arrows mark location(s) of evidence; blue color indicates low probability and red indicates high probability of passing by

5 Conclusion

This work tackled the task to explore and analyze Bluetooth tracking data visually. The question is of high interest as Bluetooth tracking is nowadays used for various pedestrian monitoring applications. The challenge related to this task is the three-dimensionality of the movement data. Thus, we constructed a three-dimensional model of the building where our experiments where 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 proposed methods were integrated and tested in an event monitoring use case. Recorded data 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. Future work needs to focus on the application of the revealed data in location based services. The application in handhelds or smartphones (e.g. for pedestrian navigation or location recommendation systems) is promising. Further work needs to focus on the temporal perspective and identification of similar mobility [4].

References

[1] G. Andrienko and N. Andrienko, A General Framework for Using Aggregation in Visual Exploration of Movement Data, The Cartographic Journal, v.47 (1), (2010), pp. 22-40

[2] G. Andrienko, N. Andrienko and S. Wrobel, Visual Analytics Tools for Analysis of Movement Data, ACM SIGKDD Explorations 9(2), 38 (2007)

[3] G. Andrienko, N. Andrienko, P. Bak, D. Keim, S. Kisilevich and S. Wrobel, A Conceptual Framework and Taxonomy of Techniques for Analyzing Movement, Journal of Visual Languages and Computing, v.22 (3), (2011), pp.213-232

[4] N. Andrienko, G. Andrienko, H. Stange, T. Liebig and D. Hecker, Visual Analytics for Understanding Spatial Situations from Episodic Movement Data, KI Journal (2012), Accepted

[5] S.G. Bottcher, C. Dethlefsen, deal: Learning Bayesian Networks with Mixed Variables (2009).URL http://www.math.aau.dk/dethlef/novo/deal. R package version 1.2-33

[6] R. Bruno and F. Delmastro, Design and Analysis of a Bluetooth-based Indoor Localization System, In Personal Wireless Communications, IFIP-TC6 8th International Conference, PWC (2003), pp. 711–725

[7] N. Friedman, I. Nachman and D. Peer, 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) (Morgan Kaufmann, 1999), pp. 206–215

[8] 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) (ACM Press, 2004), pp. 44–52

[9] Google Inc. Google earth (2010). URL http://earth.google.de/

[10] W. Hagemann and J. Weinzerl, Automatische Erfassung von Umsteigern per Bluetooth-Technologie. In: Nahverkerspraxis. (Springer-Verlag Berlin Heidelberg, Berlin Heidelberg, 2008)

[11] D. Heckerman, A Tutorial on Learning With Bayesian Networks. Tech. rep., Microsoft Research (1995)

[12] S. Leitinger, S. Gröchenig, S. Pavelka and M. Wimmer, Erfassung von Personenströmen mit der Bluetooth-Tracking-Technologie. In: Angewandte Geoinformatik 2010, 15th edn. (Addison Wesley Longman Inc., New York, 2010)

[13] A. Raffaetà, L. Leonardi, G. Marketos, G. Andrienko, N. 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, v.7 (1), (2011), pp.1-23

[14] T. Liebig and A.U. Kemloh Wagoum, Modelling Microscopic Pedestrian Mobility Using Bluetooth, In Proceedings of the International Conference on Agents and Artificial Intelligience, (SciTePress 2012), pp. 270-275

[15] T. Liebig, C. Körner and M. May, Fast Visual Trajectory Analysis Using Spatial Bayesian Networks, In ICDM Workshops (IEEE Computer Society, 2009), pp. 668–673

[16] T. Liebig, C. Körner and M. May, Scalable Sparse Bayesian Network Learning for Spatial Applications, In ICDM Workshops (IEEE Computer Society, 2008), pp. 420–425

[17] G. Marketos, E. Frentzos, I. Ntoutsi, N. Pelekis, A. Raffaeta and Y. Theodoridis, Building Real World Trajectory Warehouses, In Proceedings of the 7th International ACM SIGMOD Workshop on Data Engineering for Wireless and Mobile Access (MobiDE'08) (2008)

[18] M. Meila, 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) (Morgan Kaufmann Publishers Inc., 1999), pp. 249–257

[19] National Institute of Standards and Technology, FIPS 180-2 Secure Hash Standard, (National Institue of Standards and Technology, Washington, 2002)

[20] S. Orlando, R. Orsini, A. Raffaeta, A. Roncato and C. Silvestri, Trajectory Data Warehouses: Design and Implementation Issues, Journal of Computing Science and Engineering 1(2), 240 (2007)

[21] R Development Core Team, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2009). URL http://www.R-project.org. ISBN 3-900051-07-0

[22] H. Stange, T. Liebig, D. Hecker, G. Andrienko and N. Andrienko, Analytical Workflow of Monitoring Human Mobility in Big Event Settings using Bluetooth, In Proceedings of the 3rd International Workshop on Indoor Spatial Awareness (ACM, 2011), pp. 51–58

[23] M. Versichele, T. Neutens, M. Delafontaine and N.V. 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 32(2), 208 (2012)