Analysing spatiotemporal sequences in Bluetooth tracking data

Matthias Delafontaine[*1], Mathias Versichele[1], Tijs Neutens[1], Nico Van de Weghe[1]

[1] CartoGIS Cluster, Department of Geography, Ghent University
Krijgslaan 281 WE12
9000 Ghent (Belgium)

[*] Corresponding author: Matthias.Delafontaine@UGent.be

1. Introduction

Although existing as a communication technology since the mid-nineties, Bluetooth has only recently been employed for positioning and tracking of individuals [1-4]. Despite its limited positional accuracy, Bluetooth tracking is a low-cost alternative for true location-aware technologies. A major advantage of this technique is that it allows for the distinction of tracked subjects at the individual level. This is because Bluetooth-enabled devices broadcast a unique MAC (48-bit physical address). Furthermore, due to its widespread integration in nowadays mobile devices such as mobile phones, PDA's, laptops, headsets, etc., Bluetooth allows for unannounced tracking, i.e. tracking of subjects that are not aware of being tracked. Hence, it gives scientists the valuable potential to conduct unbiased experiments and gather uninfluenced observations of a mass of individuals.

In the large body of research on movement behaviour, considerable work has been dedicated to the extraction of patterns from motion data [5-8]. Within the abundant research concerning the analysis of sequential aspects of human activities, sequence alignment has been adopted as a promising methodology [9-12]. In line with this strand of literature, this paper will explore the potential of sequence alignment methods for the extraction of behavioural patterns within Bluetooth tracking data. In particular, the focus is on the extraction of significant clusters of subjects that share similar movement patterns as these reflect how different groups of people behave differently within the same context.

The remainder of this paper is structured as follows. Section 2 discusses the use of Bluetooth technology as a tracking system. Section 3 introduces the basic principles of sequence alignment methods. Section 4 presents an empirical case study in which sequence alignment methods are applied to analyse Bluetooth tracking data gathered at a 5-day trade fair in Ghent (Belgium). Finally, conclusions are mentioned in section 6.

2. Bluetooth Tracking System

In this paper, we consider the most basic and simple approach to employ Bluetooth technology as a tracking system. It consists of a number of Bluetooth access points (*nodes*) installed at strategic locations throughout a given study area (e.g. Figure 2). Each node continuously searches for nearby devices. Whenever a Bluetooth-enabled device enters the radio range of a node, its MAC address is logged. In this way, a dataset is obtained, consisting per node of a group of loglines of the form < *MAC*, *timestamp* >. Optionally, supplementary attributes may be logged such as the device class[1] and user-friendly name[2], although this might demand additional lookup time. As the position of the nodes is known, an approximate trajectory can be inferred for each chronological sequence of loglines sharing the same MAC.

To date, most Bluetooth tracking projects documented in the literature have relied on this concept (e.g. [1-4]). A disadvantage of the approach, however, is the limitation of the positional accuracy to the radio range of the nodes, which is a vague, rather than a crisp measure. On the other hand, apart from being robust and plain, the concept is attractive due to its easy and low-cost implementation: merely a number of Bluetooth dongles, computational units and storage units are required. Furthermore, the approach is efficient in its data collection as it does not set up true connections with devices, and thereby avoids any interaction with the individuals being tracked.

3. Sequence Alignment Methods

Having a tradition in bioinformatics for the analysis of DNA and protein strands, sequence alignment methods were first applied in social science by Abbott [10] to analyse career patterns. *Sequence alignment* is the process of equating two or more sequences using a set of eligible operations. Sequence alignment methods seek for optimal alignments by employing dynamic programming algorithms to either maximise a similarity measure, or to minimise a distance measure [11]. *Multiple alignments* (i.e. alignments of three or more sequences) are usually approximated by a procedure of multiple *pairwise alignments*, known as progressive alignment [13].

The conventional operations eligible for pairwise alignment are *identity*, *substitution*, *insertion*, and *deletion*. As they always occur together, the latter two operations are known as *indels* and are accommodated by gaps in one of both sequences. Sequences are usually represented as a string of elements consisting of one or more characters. A pairwise alignment of the character strings 'Bluetooth' and 'Blåtand'[3] is illustrated in Figure 1. To determine its optimality, the operations have to be weighted by a priori defined similarity values. The *identity* operation represents the highest similarity, whereas *substitution* is often given zero similarity and *indels* are associated with penalties (negative similarities).

---

[1] The device class is a 3-byte value that describes a device by a hierarchical classification, e.g. Phone: Cellular, Computer: Laptop.

[2] A user-friendly name is an arbitrary word or phrase most often configurable by the user.

[3] Bluetooth is named after the Danish king Harald Blåtand (940 – 981 A.D.).

B	l	u	e	t	o	o	t	h
B	l	å	–	t	a	n	d	–
✓	✓	×	–	✓	×	×	×	–

✓ identity
× substitution
– indel

Figure 1 - Alignment of 'Bluetooth' and 'Blåtand'

At least two types of analysis can be conducted on the basis of sequence alignment [12]. The most common one, which will be considered in this paper, is an analysis of clusters of similar sequences and/or representative sequences. Another possibility consists of detecting hypothetical behavioural patterns within the sequence set at hand.

4. Case study: the Horeca Expo

The aim of this case study is to apply sequence alignment methods to analyse the behavioural patterns of visitors tracked by Bluetooth. The study context, data collection, preparation and results are discussed in detail in the remainder of this section.

4.1. Data collection

The data for this case study have been collected by Bluetooth tracking at the 21st edition of the Horeca Expo (November 22-26, 2009). The Horeca Expo is the most important annual trade fair for the hotel and catering industry in Belgium. The 2009 fair has counted 53 146 visitors, most of them being professionals in the catering industry, for 607 exhibition stands. The Horeca Expo is particularly well-chosen as a setting for the examination of visiting patterns, since the daily variation and extent of additional events that may influence the temporary behaviour of visitors is strongly limited.

The Horeca Expo takes place at the Flanders Expo exhibition centre in Ghent (Belgium). The centre has eight exhibition halls over an area of about 56 000 m² (Figure 2). Each hall groups exhibition stands of a specific theme (e.g. hall 1: breweries, hall 5: kitchen contractors). 22 Bluetooth nodes, denoted $A - T$[4], have been discreetly installed throughout the entire site. The nodes are power class 2 devices which are expected to cover a radio range of about 20m, although experiments have shown that this range may vary substantially, among others due to indoor reflections. Given this presumption, it follows that on the one hand the study area is not completely covered by all nodes, and on the other hand some node pairs have an overlap in their covered areas (Figure 2). Over the entire 5-day period of the fair, 14 498 unique MACs have been observed, most of which are cell phones and the like (92%).

[4] Node *H* has been left out as it was located out of the study area.

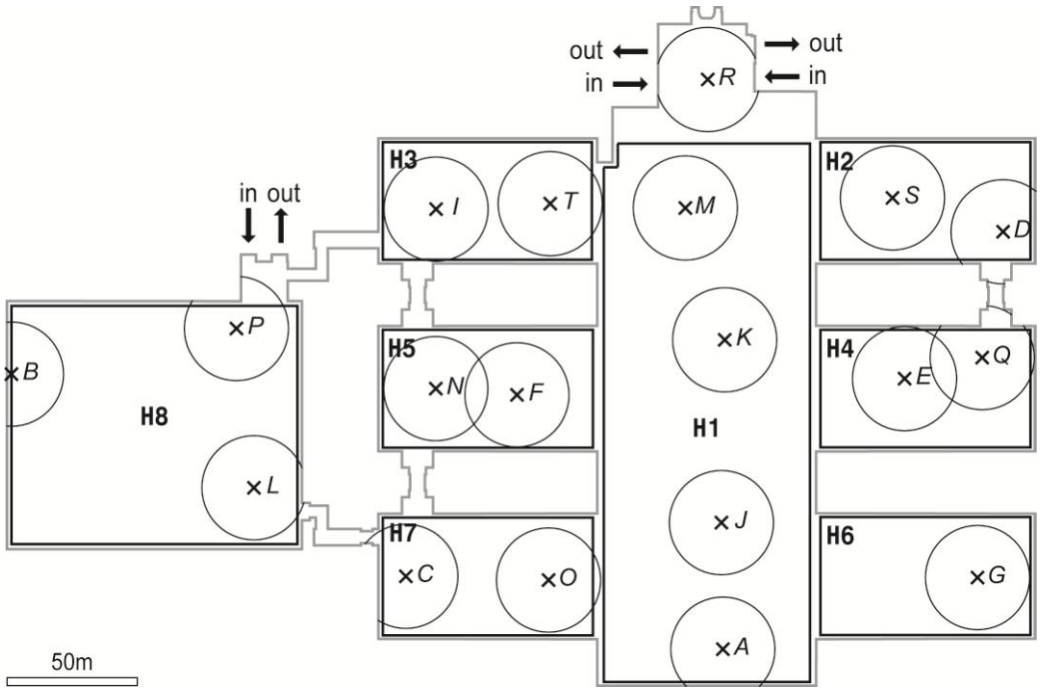

Figure 2 – Schematic map of Flanders Expo with indication of entrances and exits for visitors (arrows), exhibition halls (H1-H8, black rectangles), and Bluetooth nodes (*A-T*, x-marks) with 20m radio range (black circles)

4.2. Data preparation

Given the logline data of each Bluetooth node as described in section 2, we have determined the chronological sequence of node observations for each observed MAC. These sequences have been split up per day. To filter for noise in the data, subsequent observations by the same node that are less than one minute apart have been concatenated to one observation lasting over the entire interval. Some additional preparative steps have been taken to extract representative sequences for visitors and to exclude as much as possible the sequences of exhibitors, crew members and outlier sequences. The following restrictions have been imposed:

- The first and last observations in the sequence are observed at node *P* or *R* which are located near the visitor entrances and exits (Figure 2);
- The time span of the sequence is within one of the official opening hour intervals of the fair, i.e. each day from 10:30 a.m. to 7:00 p.m.;
- The time gaps in between two subsequent observations in the sequence have a maximum duration of 15 minutes;
- The sequence has a minimum duration of 30 minutes.
- The sequence contains observations of at least eight different nodes.

Further, the observation sequences that respect the above restrictions have been transcoded to single-character sequences to facilitate sequence alignment. To this end, a temporal unit of 3 minutes has been postulated as being the minimum episode for visiting a certain location within the fair. Hence, the observation sequences have been divided into 3-minute intervals, each of which has been allocated a character according to the following rules:

- If more than 50% of an interval is covered by observations of the same node, the node's character is allocated to the interval;
- If more than 50% of an interval is covered by observations of two nodes, the character of the node which observations cover the greater share is allocated to the interval;
- If an interval has observations of more than two nodes, a character *V* is allocated to the interval;
- In all other cases a gap character (-) is assigned.

The interpretation of sequence characters is as follows. Node characters indicate visiting events in the neighbourhood of the respective nodes; *V* characters stand for travelling episodes, i.e. visitors travelling in between two visiting events; and gaps represent visiting events at a location outside of any node's radio range. Given the above constraints and the strategic dispersion of nodes across the study area (Figure 2), it is probable that visitors remain near to the node of their last observation during gaps. As the interpretation of gaps and *V* episodes depends on their neighbouring characters, sequences consisting for more than 50% of gaps or *V* episodes have been excluded, to preserve interpretability.

4.3. Sequence alignment parameters

The area covered by a node's radio range contains multiple fair stands which hampers the analysis of visiting patterns at the stand-level. Therefore, we will rely on the thematic grouping within the exhibition halls (see 4.1) to define the mutual similarity of sequence characters. Node character episodes of nodes within the same hall can be interpreted more similar than those of nodes in different halls. An exact character match (identity) is assigned a similarity value of 10 (maximal similarity). A mismatch (substitution) is given a similarity value of 7 in the case of characters of nodes in the same hall, and 0 (maximal dissimilarity) otherwise. An exception has been made for the substitutions *A*-*K*, *J*-*M*, *A*-*M*, which received the respective similarity values of 5, 5, and 3 due to the greater distance between the corresponding nodes. Also, alternative similarity values apply for *V* characters in order to lower the priority of matching *V* episodes in the alignment process. To this end, the identity value for *V* characters is set to 4 and the substitution value with respect to all other characters to 1 (not to 0 as *V* characters are related to at least three different nodes, see 4.2).

Separate indel penalties have been considered for gap openings and for gap extensions; respectively 5 and 3.

4.4. Results

510 sequences were found to validate the restrictions imposed by the data preparation. Using the parameters specified in 4.3, a multiple alignment of these sequences has been generated within the ClustalTXY software package [11] by means of a progressive alignment procedure which consists of a pairwise alignment using a local alignment algorithm (Smith-Waterman), a neighbour-joining process, and a multiple alignment using a global alignment algorithm (Needleman-Wunsch). The neighbour-joining process has produced a guide tree

which totals 509 hierarchical clusters (Figure 3). The clusters observed in this guide tree may assist in the determination of a typology of different visitor behavioural patterns. The number of members in a cluster can thus be considered an indicator for the importance of the corresponding behavioural pattern. Sequences in smaller clusters, however, tend to have a more elements in common.

It is usually considered up to the analyst to determine the number of clusters. Three major clusters (1-3) can be observed at the top of the hierarchy (Figure 3). At this level, the aligned sequences hardly share common characteristics, if at all. To enable visual exploration, node characters in the alignment have been colour coded to the hall where they are located. On the basis of visual supervision, an exhaustive subdivision has been made into 21 non-overlapping subclusters (1.1-3.8) (Figure 3). For each of these subclusters the number of members, the main pattern(s), and the average and median sequences have been listed in Table 1 and Table 2. The average and median sequences are representative sequences of the cluster [11]. In analogy to their homonymous summary statistics, these sequences respectively minimise the sum of squared distances and the sum of distances to all other members of the cluster.

The results in Table 1 and Table 2 reveal some interesting aspects about the behaviour of Horeca Expo visitors. It can be observed that visitors tend to spend more time at the entrance than at the exit, which can be explained by typical entrance activities such as registering, informing and depositing luggage in a cloakroom. Inferences can also be made concerning the attractiveness of locations, although these might be misleading as not all halls have been equally covered by Bluetooth nodes. The abundant hall 1 episodes - especially within cluster 3 – reflect that the main hall is also the most important one in terms of visits, as could be expected given its size and central location. Other inferences concern the chronology of hall combinations. For instance, the concatenation of hall 2 and hall 4 episodes, or more specifically D and Q episodes, is common in many clusters, which underlines the importance of the direct passage that connects both halls (Figure 2).

Notwithstanding the spatiotemporal structure found within the visitor sequences, evidence of strong and highly similar behavioural patterns is lacking. Conversely, the high number of subclusters yet considered and the limited similarity among the members within some clusters, demonstrate that there exist many diverse behavioural patterns among visitors at the Horeca Expo, both in terms of duration and the order of visiting locations.

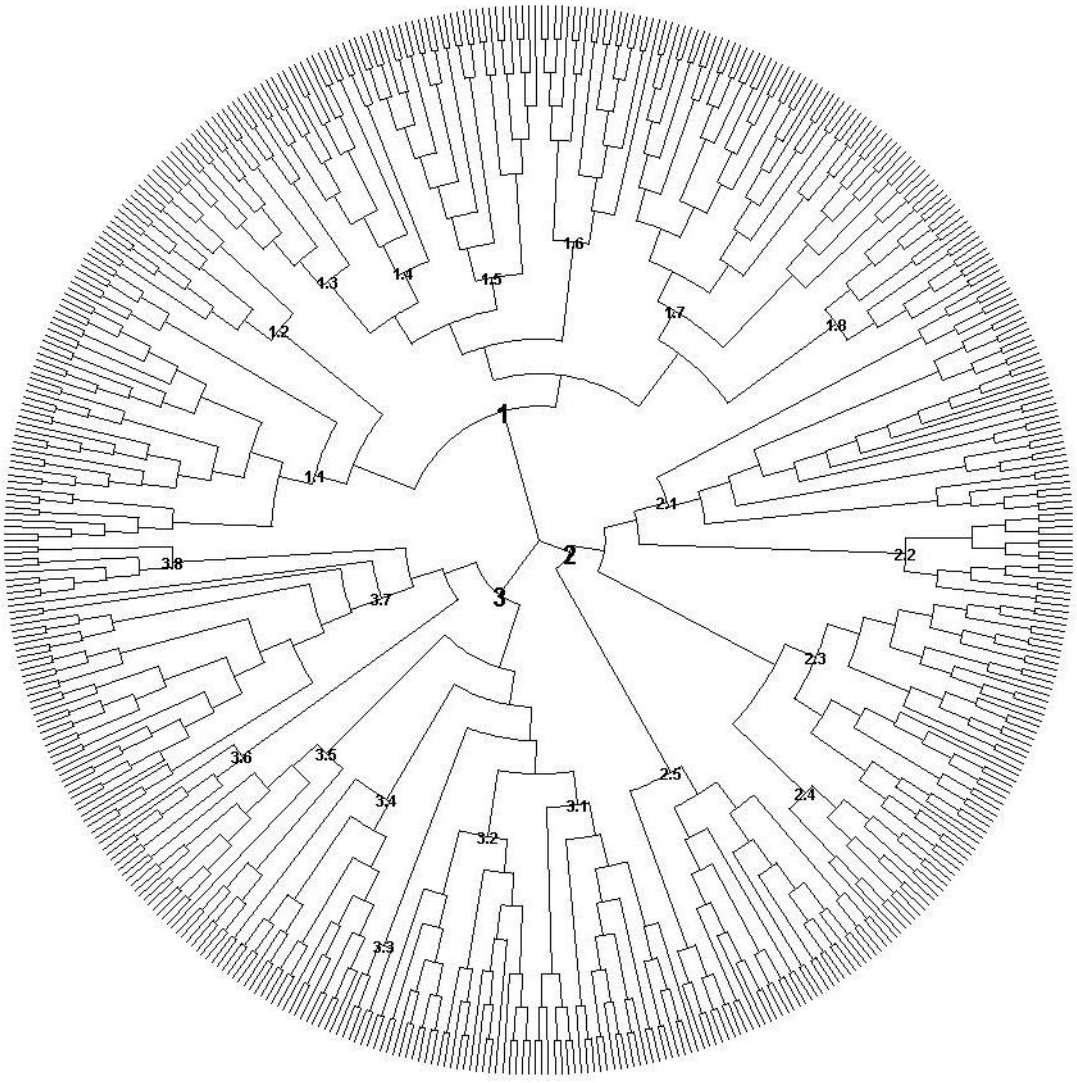

Figure 3 – Multiple alignment guide tree with clusters and subclusters labeled at their root node

Cluster	Members	Common patterns
1.1	47	… - **H8** - **H7** - **H1** - H2|H4 - …
1.2	17	… - **H1** - H8 - … - H1
1.3	17	… - H1 - **H4** - …
1.4	18	… - H1|H5|H3 - **H7** - … - H1
1.5	29	H3|H8 - H7 - H1(A) - …
1.6	19	… - H1(A|J) - … - **H1(A|J)** - H7 - …
1.7	49	H7 - H1 - H6 - H4 - H2 - H1
1.8	17	H1|H2|H8 – H3 – H4 – H1 - …
2.1	38	…- H1|H7 - **H8(B)** – H1|H7 …
2.2	15	**H2** - **H4** - **H6** - **H1** - H7 - H5 - H8 - H3 - H1
2.3	44	…|H1 – **H2** – H4|H6 – H1 - …
2.4	14	… - **H6** – H1|H7|H3
2.5	34	H8 – H1 – H2 – H4 - **H5** – H1 - …
3.1	26	H7|H8|H5 - H4 - **H1(M)**
3.2	26	**H1(M)** - H2|H4|H6|H8 - H1
3.3	5	**H1(M)** – H3|H5|H7 – H6 – H4 – H2
3.4	22	… - H1 – **H3(T)** – H1 - …
3.5	19	… - **H1(K)** – H3|H5 - …
3.6	7	**H1** – H2 – H4 – H6|H1 – H7 – H5 – H3 - …
3.7	38	… - H1(J) – H2|H4|H7 - …
3.8	9	H2 – H4 – **H1(J)** - … - **H3** - …

Pattern legend	
H1	episode at hall 1
H1(A)	episode at hall 1, node A
H1|H2	episode at hall 1 and/or hall2
…	episode at one or more halls
H	predominant episode
H	frequent episode
H	occasional episode

Table 1 – Number of members and common patterns per cluster

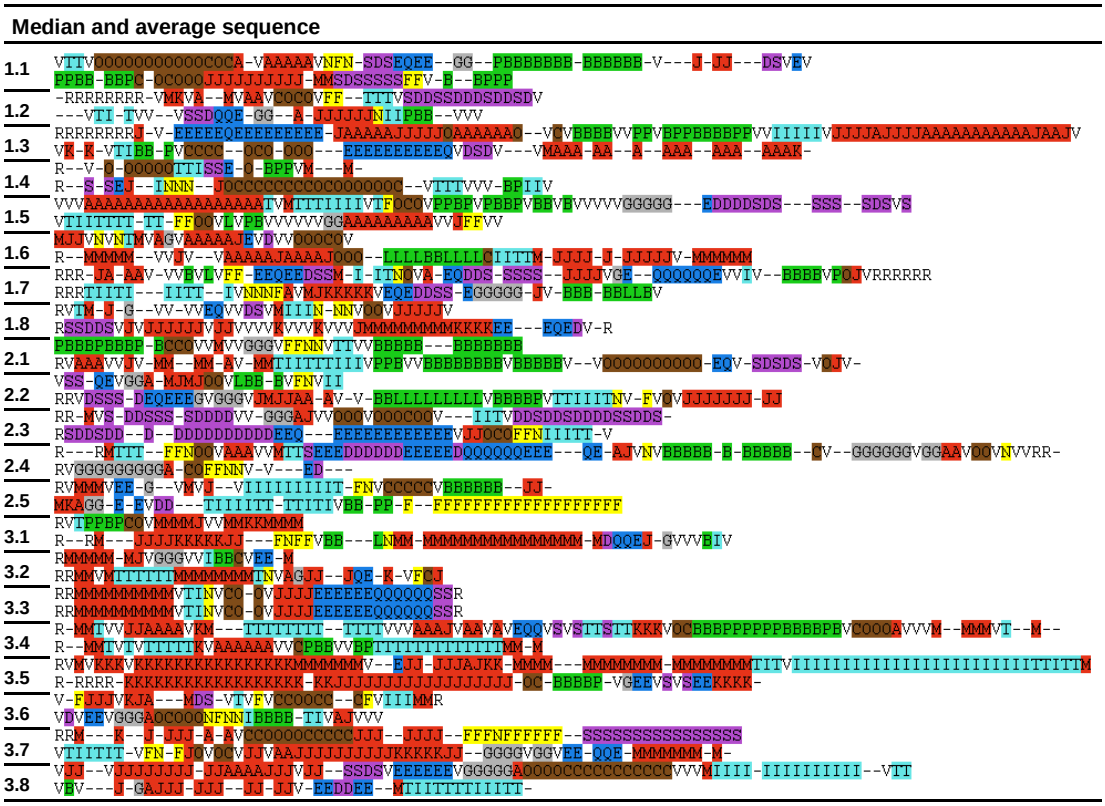

Table 2 - Median (top) and average (bottom) sequence per cluster (colour legend: hall 1, hall 2, hall 3, hall 4, hall 5, hall 6, hall 7, hall 8)

5. Conclusions

In this paper, we have explored the potential of sequence alignment methods to analyse Bluetooth tracking data. After a brief introduction on Bluetooth tracking and sequence alignment methods, a case study has been presented on the analysis of spatiotemporal behavioural patterns of visitors tracked by Bluetooth nodes at a big trade fair. We have shown that sequence alignment methods can be successfully adopted to analyse Bluetooth tracking sequences. The results have revealed some important and plausible insights about the visiting patterns at the fair. They confirm the ability of a simple Bluetooth tracking system to capture the spatiotemporal behaviour of large crowds of individuals at a mass event.

Despite the above contributions, some aspects still limit the potential of sequence alignment methods for the analysis of tracking data. Unlike the structure of nucleotides in a strand of DNA, spatiotemporal sequences within tracking data might differ very much amongst tracked individuals, both with respect to sequence composition as with respect to the number of elements (duration). In sequence alignment, the latter aspect may cause a large number of gaps, for which there is yet no consensus on their interpretation [13]. Shoval and Isaacson [12] recognize the lack of a solid method to assess the reliability of alignments, as well as the lack of knowledge on the impact of the spatial and temporal scale on the results. Also, there is no consensus method or standard calibration procedure for the setting of sequence alignment parameters such as indel penalties. Future progress on these issues will support stronger and more refined interpretations of alignment results.

Acknowledgments

The authors express their great gratitude to Artexis for allowing and supporting Bluetooth tracking at the Horeca Expo. The Research Foundation – Flanders is gratefully acknowledged for funding the research of Matthias Delafontaine and Tijs Neutens; the agency for Innovation by Science and Technology for funding the research of Mathias Versichele.

References

1. Van Londersele, B., Delafontaine, M., Van de Weghe, N.: Bluetooth Tracking - a spy in your pocket. GIM International. Geomares Publishing (2009) 23-25
2. Fallast, K., Scholz, A., Ekam, H.W.: Sichere Basis für Verkehrsplanung: Erfassung von Fahrgastströmen via Bluetooth. Der Nahverkehr (2008) 72-75
3. Wasson, J.S., Sturdevant, J.R., Bullock, D.M.: Real-Time Travel Time Estimates Using Media Access Control Address Matching. ITE Journal **June 2008** (2008) 20-23
4. Bullock, D.M., Haseman, R., Wasson, J.S., Spitler, R.: Anonymous Bluetooth Probes for Measuring Airport Security Screening Passage Time: The Indianapolis Pilot Deployment. Transportation Research Board Annual Meeting 2010 (2010) Paper #10-1438
5. Dodge, S., Weibel, R., Lautenschutz, A.: Towards a taxonomy of movement patterns. Inf Visualization **7** (2008) 240-252
6. Laube, P., Imfeld, S., Weibel, R.: Discovering relative motion patterns in groups of moving point objects. International Journal of Geographical Information Science **19** (2005) 639 - 668
7. Gudmundsson, J., van Kreveld, M., Speckmann, B.: Efficient Detection of Patterns in 2D Trajectories of Moving Points. Geoinformatica **11** (2007) 195-215
8. Gottfried, B., Aghajan, H.: Behaviour Monitoring and Interpretation - BMI: Smart Environments, Volume 3 Ambient Intelligence and Smart Environments. IOS Press (2009)
9. Joh, C.-H., Arentze, T., Hofman, F., Timmermans, H.: Activity pattern similarity: a multidimensional sequence alignment method. Transportation Research Part B: Methodological **36** (2002) 385-403
10. Abbott, A.: Sequence Analysis: New Methods for Old Ideas. Annual Review of Sociology **21** (1995) 93-113
11. Wilson, C.: Activity patterns in space and time: calculating representative Hagerstrand trajectories. Transportation **35** (2008) 485-499
12. Shoval, N., Isaacson, M.: Sequence Alignment as a Method for Human Activity Analysis in Space and Time. Annals of the Association of American Geographers **97** (2007) 282-297
13. Wilson, C.: Reliability of sequence-alignment analysis of social processes: Monte Carlo tests of ClustalG software. Environment and Planning A **38** (2006) 187-204