Master's thesis GEO 511

# Analysing spatio-temporal behavior of shopping mall visitors using sequence alignment methods

Submitted by: Simon Jakob (10-711-000)

Date of Submission: June 30, 2016

**Supervisors:** Dr. Martin Tomko & Prof. Dr. Robert Weibel

**Faculty Member:** Prof. Dr. Robert Weibel

**Geographic Information Science (GIS)**

Department of Geography

University of Zurich

Winterthurerstrasse 190

8057 Zurich - Switzerland

# Contact

## Author

**Simon Jakob**
Ohmstrasse 28
8050 Zürich – Switzerland
simon.jakob@sunrise.ch

## Supervisors

**Dr. Martin Tomko**
Department of Infrastructure Engineering
University of Melbourne
Parkville VIC 3010
Melbourne – Australia
tomkom@unimelb.edu.au

**Prof. Dr. Robert Weibel**
Geographic Information Science (GIS)
Department of Geography
University of Zurich
Winterthurerstrasse 190
8057 Zurich – Switzerland
robert.weibel@geo.uzh.ch

# ACKNOWLEDGEMENTS

For their support during the completion of this Master's thesis I would like to address special thanks to:

- Dr. Martin Tomko for productive discussions, constructive critique and the occasional wake-up call not depending on the meeting taking place on campus or via Skype Zurich-Melbourne.
- Prof. Dr. Robert Weibel for guiding me to the research field of my interest and managing my thesis from a distance.
- Julia Gross for proof reading and continuous support.

Thank you.

With the end of my studies coming closer, I feel grateful towards the Department of Geography of the University of Zurich for making this such a fun and instructive time, towards IBV Hüsler AG for hiring me during my Master's and giving me the much needed change from every day life as a student and towards endurance sports when I needed to get my head off both work and studies. This all would not have been nearly as enjoyable without the continuous support of my family and Julia Gross being always there for me, thank you all. I've had 6 great years and am curious what's about to come next!

Simon Jakob

June 2016

Acknowledgements

# ABSTRACT

Wi-Fi networks, implemented in many indoor environments such as shopping malls or airports, can be used to track the position of a device within the network. This can be used to the advantage of the user through location based services, but it also opens possibilities for the scientific analysis of the revealed movement traces through indoor environment that could by other means (e.g. GPS) not be observed. This thesis analyzes such a Wi-Fi tracking dataset covering approx. 260'000 trajectories of approx. 120'000 unique users of a big shopping mall in Sydney, Australia, that the TRIIBE project of the RMIT and the University of Melbourne was granted access to.

By pre-processing the dataset in an adequate way and by then applying well-calibrated movement mining tools, this thesis strives to identify movement patterns in the data, such as recurrent patterns of behavior or similar groups of people moving through the shopping mall in a similar manner. As for the analysis of movement, like for all geographic phenomena, scale matters, a cross scale analysis, investigating the influence of spatial and temporal scale on the found patterns, is performed.

As the dataset, due to its origin, features relatively low spatial and temporal resolution, basic movement mining methods are not expected to show good performance, which is why the more abstract sequence alignment method (SAM) was applied. SAM was developed by microbiologists to analyze DNA-sequences and introduced into spatial studies by Bargeman et al. (2002) basically analyzing sequences of locations. This thesis for the first time applies SAM to a Wi-Fi tracking dataset.

The found patterns were evaluated using statistical tests and cluster validation measures. Recurrent patterns of behavior could be detected, as could be shown that user categorization based on some temporal characteristics, such as return period or usual time of day of the shopping trip, showed a significant relation to clusters of similar trajectories. Other temporal characteristics, such as usual day of week of the shopping trip, did not show such a relationship. The spatial granularity of the data was not observed to have a strong influence on the detected patterns and relations and the same counts, with some methodological limitations, also for the temporal granularity. Finally, SAM was evaluated to be an appropriate movement mining tool for such sparsely sampled movement data, despite limits regarding evaluation and performance.

Abstract

# ZUSAMMENFASSUNG

Wi-Fi Netzwerke, wie sie in vielen öffentlichen Gebäuden wie Einkaufszentren oder Flughäfen zu finden sind, können benützt werden um die Position eines Gerätes innerhalb des Netzwerkes zu bestimmen. Dies kann durch sogenannte location based services zum Vorteil der Benutzer verwendet werden, aber es ermöglicht auch die wissenschaftliche Analyse von Personenbewegungen im Innern von Gebäuden, welche mit anderen Techniken (z.b. GPS) nicht beobachtet werden könnten. Diese Masterarbeit analysiert einen solchen Wi-Fi tracking Datensatz, der ca. 260’000 Trajektorien von 120'000 verschiedenen Besuchern eines Einkaufszentrums in Sydney, Australien umfasst. Der Datensatz wurde dem TRRIBE-Projekt des RMIT und der Universität Melbourne zur Verfügung gestellt.

Diese Masterarbeit versucht, durch eine adäquate Vorbereitung der Daten und den nachfolgenden Einsatz von movement mining Methoden, Bewegungsmuster wie z.b. repetitives Verhalten oder ähnliches Bewegungsverhalten von ähnlichen Kundengruppen zu identifizieren. Da für die Analyse von Bewegungen, wie für die Analyse aller andern geographischen Phänomene, der Betrachtungsmassstab eine wichtige Rolle spielt, wird der Einfluss von verschiedenen räumlichen und zeitlichen Massstabsebenen auf die gefundenen Bewegungsmuster anhand einer cross-scale Analyse geprüft.

Aufgrund seiner Herkunft weist der Datensatz eine relativ tiefe räumliche und zeitliche Auflösung auf, was zur Annahme führt, dass die Anwendung von traditionellen movement mining Methoden nicht zu guten Resultaten führt. Aus diesem Grund wurden sequence alignment Methoden (SAM) angewendet. SAM wurde von Mikrobiolgen entwickelt, um DNA-sequenzen zu analysieren und wurde von Bargeman et al. (2002) ins räumliche Forschungsgebiet eingeführt, um Sequenzen von räumlichen Positionen zu analysieren. Diese Masterarbeit wendet SAM zum ersten Mal in Kombination mit einem Wi-Fi tracking Datensatz an.

Die gefundenen Bewegungsmuster wurden mithilfe von statistischen Tests und Cluster-Validierungskennzahlen evaluiert. Repetitives Verhalten konnte entdeckt werden, wie auch gezeigt werden konnte das, bezüglich ihrer zeitlichen Charakteristiken wie z.b. ihr Wiederkehrverhalten oder ihre normale Shopping-Uhrzeit, ähnliche Gruppen von Kunden eine signifikante Beziehung zu Cluster ähnlicher Trajetorien aufweisen. Andere zeitliche Charakteristiken wie z.b. häufigster Einkaufstag zeigten keine solche Beziehung. Ausserdem wurde die räumliche Auflösung der Daten als nicht kritisch für die identifizierten Bewegungsmuster bewertet, was, mit methodischen Einschränkungen, auch auf die zeitliche Auflösung zutrifft. Zum Schluss wurde SAM als angebrachte Methode für einen solchen Datensatz bewertet, obschon Einschränkungen bezüglich Evaluation und Performanz der Methode aufgezeigt wurden.

# CONTENTS

Contents

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF EQUATIONS

# LIST OF ABBREVIATIONS

| | |
|---|---|
| **AD** | Average Distance |
| **ADM** | Average Distance between Means |
| **AGNES** | AGglomerative NESting |
| **AP** | Access Point |
| **APN** | Average Portion of Non-overlap |
| **CLARA** | Clustering LARge Applications |
| **DIANA** | DIvisive ANAlysis |
| **FANNY** | Fuzzy ANalYsis |
| **FOM** | Figure Of Merit |
| **GEP** | Gap Expanding Penalty |
| **GOP** | Gap Opening Penalty |
| **GPS** | Global Positioning System |
| **LBS** | Location Based Service |
| **LCS** | Longest Common Subsequence |
| **MAC** | Media Access Control |
| **OM** | Optimal Matching |
| **PAM** | Partitioning Around Medoids |
| **RSSI** | Received Signal Strength Indicator |
| **SAM** | Sequence Alignment Methods |
| **SOM** | Self-Organizing Maps |
| **SOTA** | Self-Organizing Tree Algorithm |
| **TraMineR** | Trajectory MineR |
| **TRIIBE** | TRacking Indoor Information BEhaviour |
| **WI-FI** | Used as synonym for WLAN |
| **WLAN** | Wireless Local Area Network |

# 1 INTRODUCTION

## 1.1 CONTEXT

People's movement traces provide valuable insight into their behavior (Tomko et al. 2014). Movement mining, the research field analyzing these traces is defined as *"…[aiming] for conceptualizing and detecting non-random properties and relationships in movement data that are valid, novel, useful, and ultimately understandable*" (Laube 2014, p. 31). It can therefore result in knowledge about how people act in time and space in different environments.

This thesis is part of the TRIIBE (Tracking Indoor Information BEhavior) project of RMIT and the University of Melbourne. The research group was granted access to an extensive Wi-Fi-tracking dataset from a shopping center in Sidney, Australia. This dataset can be analyzed in respect to web behavior and information needs (Ren et al. 2015), but it can also be seen as a movement trace of the shoppers through the mall. This thesis evaluates movement mining tools for Eulerian trajectories, such as those sensed by Wi-Fi tracking in indoor environments. As the dataset, due to its Wi-Fi tracking origin, comes with limitations regarding spatial and temporal resolution, this thesis tests the applicability of an elaborate movement mining method, namely sequence alignment methods (SAM).

SAM, first used in the 1970s in microbiology (Mount 2004), was introduced into the social sciences by Abbott & Forrest (1986) and the first application in a spatial study was performed by Bargeman et al. (2002). As the method's aim is to assess the similarity of sequences and as movement traces can be modeled as a sequence of locations, SAM can be used as a movement mining tool, which was demonstrated by several studies in the last ten years (Wilson 2008; Shoval & Isaacson 2007; Delafontaine et al. 2012; De Groeve et al. 2015). The findings of this analysis may ultimately be used to gain insight into the indoor movement behavior of the shoppers.

## 1.2 RESEARCH AIMS, RESEARCH QUESTIONS AND HYPOTHESES

The application of SAM to the Wi-Fi tracking dataset of a big shopping mall performed in this thesis aims at revealing new insights into the indoor movement behavior of the shoppers, but also wants to add new facets to the research field of spatial applications of SAM. Specifically, the existence of recurrent patterns of behavior of shoppers as suggested in the shopping and marketing literature (Sheth & Park 1974; Laaksonen 1993; Mulligan 1987) is tested (RQ1), as is the existence of similar behavior of homogenous groups as suggested by (Zeithaml 1985; Wesley et al. 2006)(RQ2). A contribution to the SAM research field is made by thoroughly calibrating the algorithms and by analyzing the influence of different spatial and temporal scale levels on the results of the movement mining (RQ3).

- **RQ1**: To what extent can spatial patterns of recurrent movements in indoor environments be detected?
  - **Hypothesis 1**: Symbolic representation of the trajectories can be applied in combination with sequence alignment methods (SAM) to identify recurrent movement patterns in Eulerian datasets of indoor movement.
- **RQ2**: What is the relation between clusters of similar trajectories and types of visitors?

  - **Hypothesis 2**: The type of shoppers has a significant relationship with the trajectory clusters. Shoppers with similar characteristics (e.g. usual time of day of their visits) are expected to have similar trajectories.

- **RQ3**: How does spatial and temporal resolution affect the results computed to answer RQ1 and RQ2?

  - **Hypothesis 3**: Trajectories expressed at coarser spatial and temporal granularities are less specific and discriminating than if analyzed at finer granularities. A granularity as fine as possible produces best results for RQ1 and RQ 2.

## 1.3 THESIS STRUCTURE

After this introduction, an overview over the related work and research gaps is presented in Chapter 2. Chapter 3 characterizes the dataset and explains the pre-processing steps performed. Chapter 4 discusses the implementation of the central methods of the thesis: SAM and clustering. In Chapter 5 the results of the analysis are presented, which are discussed in Chapter 6. Finally, Chapter 7 summarizes the main findings and contributions of the thesis and gives a short outlook on future work.

# 2 RELATED WORK

In this chapter related work from the research fields relevant to this thesis is discussed. Section 2.1 introduces the data collection technique of Wi-Fi tracking, used to acquire the dataset analyzed in this thesis. Section 2.2 gives a broad overview of movement mining, which is the research field this thesis strives to contribute to. Lastly, Section 2.3 presents movement mining studies using sequence alignment in more depth, as sequence alignment is the methodology applied in the analysis part of this thesis.

## 2.1 WI-FI TRACKING

Knowing the absolute and relative position of a device and its associated user in space introduces a multitude of scientific and commercial possibilities. Historically GPS positioning has been the technology of choice to conduct mobile phone positioning (Zandbergen 2009). However, the expansion of wireless local area networks (WLANs) increased the interest into positioning systems using this infrastructure (Manodham et al. 2007). Studies in well-controlled indoor environments (Bell et al. 2010; Manodham et al. 2007; Mok & Retscher 2007) showed promising results, implying that the technique may very well be used to supplement GPS positioning where GPS is unreliable, as for example indoors, in urban canyons or in natural environments with extreme relief (Bell et al. 2010).

In this related work section, the different Wi-Fi tracking methods are presented (Section 2.1.1.), possible applications of the resulting positions and location logs are discussed (Section 2.1.2), and privacy issues are addressed (Section 2.1.3).

### 2.1.1 Positioning methods

Wi-Fi positioning uses the received signal strength indicator (RSSI), a measure of the strength of the signal of an access point (AP, WI-FI hotspot) at the device, to deduct a relative distance of the device to this AP (Mok & Retscher 2007). Because of signal attenuation by the physical properties of the environment (passage through air etc.) the stronger the RSSI is for a given AP, the closer the device is assumed to be to this AP. Wi-Fi tracking can be implemented in-device or in-network and uses techniques such as fingerprinting, trilateration and the cell of origin methodology (Bell et al. 2010).

#### *2.1.1.1 In-device vs. in-network*

Traditional Wi-Fi tracking happens in-device, meaning that some kind of software or app storing the needed information has to be installed on the tracked device (Bell et al. 2010). In-network tracking in contrast describes a methodology for which no information from the device is needed, all the information can be retrieved from the network (Lin et al. 2006).

The advantage of in-network compared to in-device tracking is the low instrumentalisation, which leads to bigger populations that can be tracked. However, regarding positioning technique (see next section) the methods using more than one AP connection at a time are more complex and expensive to implement in-network than in-device, leading to many in-network tracking datasets only featuring one AP connection at a time (Bai et al. 2014).

#### *2.1.1.2 Positioning techniques*

Wi-Fi tracking positioning techniques can be categorized into methods that use multiple AP connections at a time (fingerprinting, trilateration) and methods that use only one AP connection at a time (cell of origin).

Fingerprinting consists of an offline and an online phase. During the offline phase, the system is calibrated with RSSI-signatures, a set of RSSI-values from different APs, recorded at different positions throughout the observation perimeter. During the online phase, a fingerprinting algorithm compares the signal signature a device detects to the previously acquired database and assigns the location of the most similar fingerprint in the database to the device. (Bell et al. 2010)

Trilateration directly converts RSSI-values to distances and when at least three distances to APs are known, the position of the device can be inferred mathematically, similar to GPS-positioning (Figure 1). However, the signal strength-to-distance function can not be assumed to be linear nor continuous in environments susceptible to radio interference and multipath effect as a consequence of walls, floors etc. The determination of an appropriate signal strength-to-distance function therefore requires careful calibration. (Bell et al. 2010)

*Figure 1: Distance from three APs calculated using RSSI values (Source:* (Bell et al. 2010)*)*

Fingerprinting and trilateration are both reported to achieve a positioning accuracy of 1-3 m (Mok & Retscher 2007).

If only one connection with one AP at a time is recorded, traditional Wi-Fi tracking approaches (trilateration, fingerprinting) cannot be applied (Bai et al. 2014). Bai et al. (2014) therefore developed a positioning methodology that only needs one single Wi-Fi connection. It extends the cell of origin methodology and consists of two phases: the cell determination phase and the user tracking phase.

In the cell determination phase, Voronoi polygons are created around each AP (Figure 2). These polygons can then be manually adjusted based on the physical surrounding (Figure 3). (Bai et al. 2014)

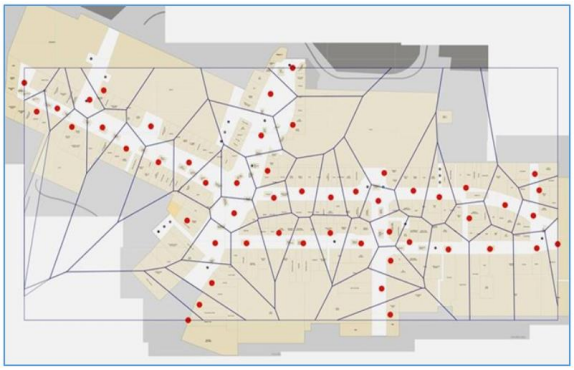


*Figure 2: Voronoi polygons around APs (Source: Bai et al. 2014)*

*Figure 3: Adjusted Voronoi polygons respecting physical layout of the environment (Source: Bai et al. 2014)*

In the user location tracking phase, the user is assigned to the cell of the AP the user is currently connected with (Bai et al. 2014).

Bai et al. (2014) achieved 96% correct cell assignments and in the remaining 4% the user was close to the cell boarder. The positional accuracy of this approach is dependent on the density of APs but can in natural environments not be expected to get close to the positional accuracy of the presented in-device tracking approach.

### 2.1.2 Applications

The possible applications of the Wi-Fi tracking datasets are manifold and can be summarized with the term *location based services (LBS).* LBS include tasks such as navigation assistance, commercial services, recreation, tracking and emergency services (Zandbergen 2009). Mok and Retscher (2007) add fleet management and location identification to the list whereas Rekimoto et al. (2007) concentrate more on user location logs and name applications such as activity pattern visualizations, life pattern arithmetic or event detection. Furthermore, Woo et al. (2011) showed that Wi-Fi positioning is not only interesting for tracking humans but also for tracked machines by investigating the possibilities of tracking labor resources within a construction site.

Wi-Fi tracking datasets can also be used for knowledge discovery, such as extracting information about movement behavior of tracked entities, which is also the aim of this thesis. How movement data can be used to this end is discussed in Section 2.2, in which also a number of studies using datasets similar to Wi-Fi tracking datasets are presented.

### 2.1.3 Privacy issues

Mobile phones with built-in cameras, microphones and position awareness may be of use to the user when he/she studies his/her mobility or health, when he/she tries to understand his/her exercise habits or the frequency of his/her interaction with family and friends or when he/she uses LBS to find the closest restaurant or to navigate through unknown territory (Cheng et al. 2006; Shilton 2009). On the other hand, all these data can also reveal regular locations, habits and routines to whoever collects, retrieves or purchases the data (Shilton 2009). It is feared that government agencies and private companies use this kind of data, which makes its protection important. Regarding position data however, protection measures such as *cloaking*, masking the user's precise location, often interfere with the quality of the requested LBS, as was advocated by Duckham & Kulik (2005) by introducing

their model of obfuscation. This may be one of the reasons why *"the collection of large-scale, longitudinal data about human mobility is now commonplace"* (Sapiezynski et al. 2015, p.1).

Wi-Fi logs typically only allow location tracking if one knows the location of all the APs. Sapiezynski et al. (2015) however showed that it is possible to extract location logs from Wi-Fi logs by combining them with sparse GPS location fixes, a method known as *wardriving*. As Wi-Fi logs potentially contain other sensitive information, such as online behavior (e.g. Ren et al. 2015), scientific work with such data requires utmost carefulness not to violate personal privacy of the study participants.

## 2.2 MOVEMENT MINING

Movement mining, already in the introduction defined as *"…[aiming] for conceptualizing and detecting non-random properties and relationships in movement data that are valid, novel, useful, and ultimately understandable*" (Laube 2014, p. 31), is a form of data mining in movement databases (Laube 2014). Data mining itself is the core step in the process of knowledge discovery in databases as described by Fayyad et al. (1996) (Figure 4). A selection of the original data gets preprocessed and transformed, to then apply *"specific algorithms to extract patterns from data"* (Fayyad et al. 1996, p. 39), known as data mining, and in a last step interpret and evaluate these patterns to produce knowledge.

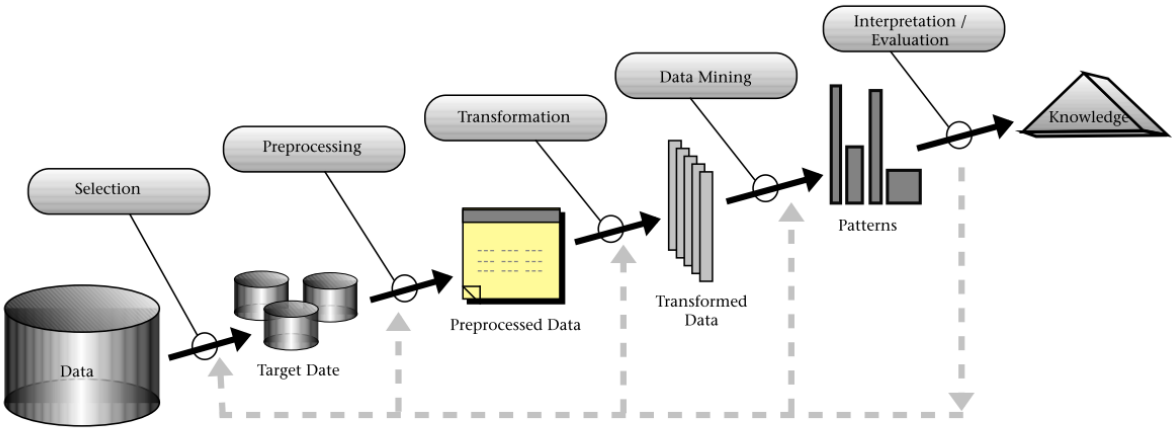



*Figure 4: The five steps of knowledge discovery in databases. Source:* Fayyad et al. (1996).

In this section, first different conceptual models capturing different aspect of movement data are presented (Section 2.2.1), then data uncertainty in movement data is discussed (Section 2.2.2), subsequently four different movement mining tasks are distinguished, of which similarity and clustering, due to their importance for this thesis, are discussed in most depth (Section 2.2.3), and lastly evaluation strategies for movement mining results are presented (Section 2.2.4). This Section heavily draws from Laube (2014) which gives an extensive, well organized and easy to understand overview of the field and therefore is where to look for more detail, in the case this section should lack it in some places.

### 2.2.1 Conceptual models for movement traces

The choice of the conceptual model of movement strongly affects the analysis methods one can apply (Laube 2014). As it is not strictly data-inherent, the choice of an appropriate conceptual model is of

high importance (Laube 2014). The different models can be distinguished in three dimensions: langrangian/Eulerian movement, constrained/unconstrained movement and discrete/continuous movement spaces.

Lagrangian movement describes absolute changes in a moving object's location whereas Eulerian movement considers *"changes in location relative to known, fixed points in space"* (Both et al. 2012, p. 3). Figure 5 shows the same movement, once captured by sampling it in a here constant time interval of 5 units and thereby using the Lagrangian conceptual model, and once captured as it passes nodes and checkpoints and thereby using the Eulerian conceptual model.

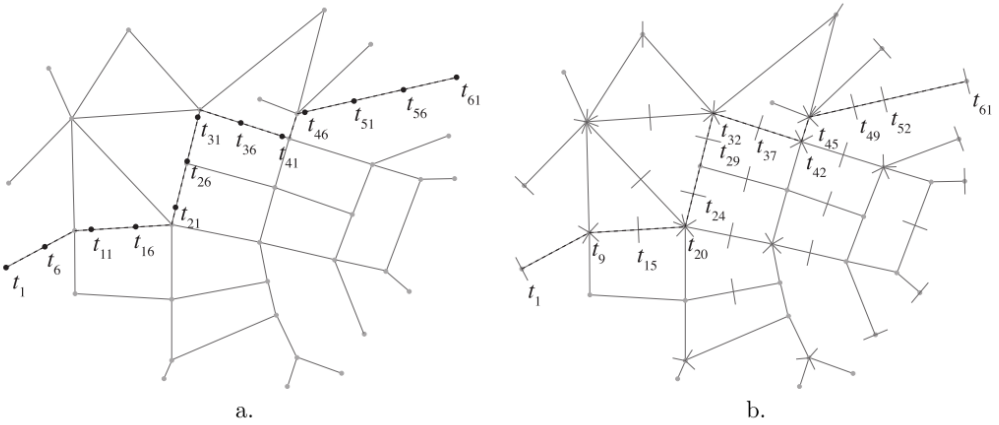



*Figure 5: Movement of an object as a sequence of timestamped locations (a. Lagrangian movement) and passed checkpoints (b. Eulerian movement). Source:* (Both et al. 2012)*.*

It cannot always be assumed that movement happens in a totally unconstrained space such as a bird's flight, as it can also be limited by some kind of boundaries (e.g. a body of water for a fish), which brings us to the differentiation of constrained and unconstrained movement environments. As human movement is most often linked to some kind of traffic infrastructure (e.g. street network for cars, building architecture of a shopping center) it is constrained in its dispersion, a fact that can be important to consider when modeling and analyzing movement data (Laube 2014).

A last dimension proposed by Laube (2014) is the differentiation between continuous movement spaces, e.g. the movement of a cow on a meadow, and discrete movement spaces, e.g. the movement of a visitor of a shopping center, captured as a sequence of Wi-Fi hotspot areas. The imperative to use discrete movement spaces often is dictated by the data source, but can also be chosen deliberately, e.g. to reduce redundancy and data set size of GPS-logs (Richter et al. 2012). Figure 6 depicts how a thus compressed view of movement shows the information more approachable and thereby facilitates subsequent analysis steps.

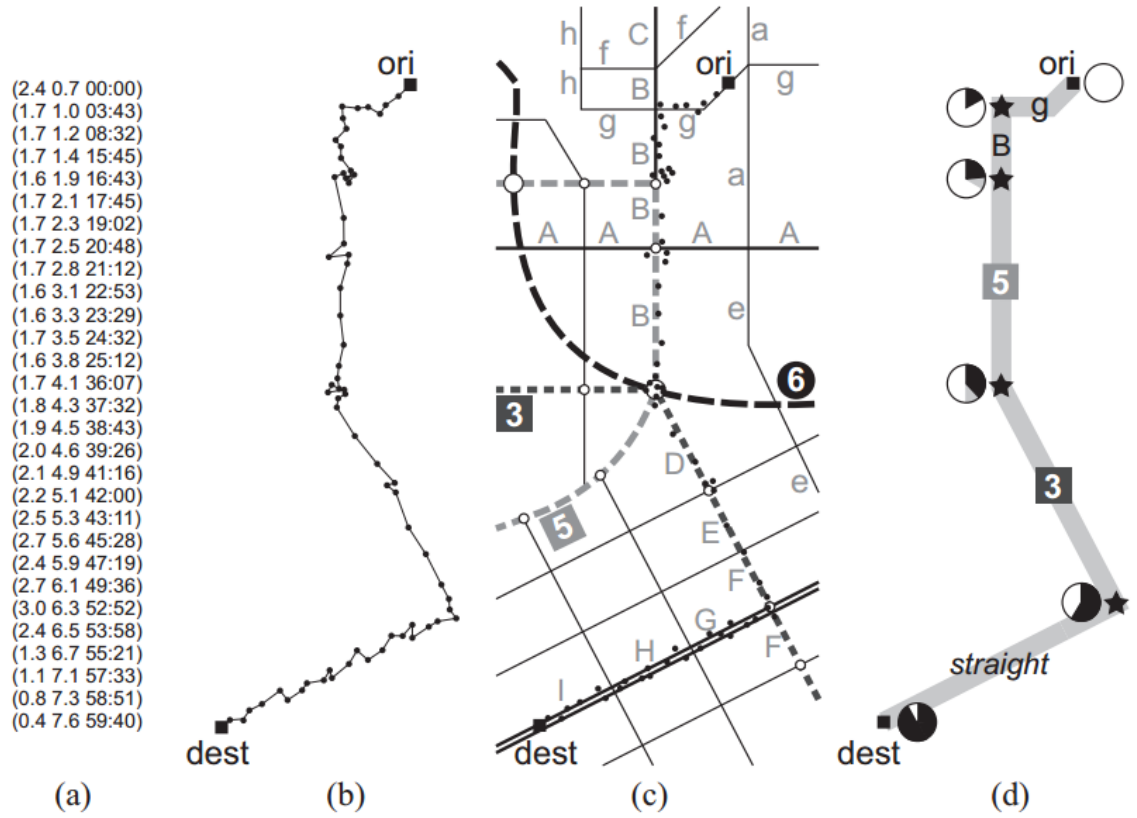



*Figure 6: Movement as raw positional data (a), two-dimensional trajectories without (b) and with (c) semantical information, and compressed to discrete movement space (d). Source:* Richter et al. (2012).

The differentiation of these dimensions of conceptual movement models by Laube (2014) is by no means uncontested. Andrienko et al. (2008) for example propose another categorization of movement data as consisting of either time-based recordings (position is stored at regular intervals), change-based recordings (a record is made when an object moves), location-based recordings (record is made when an object comes close to a location e.g. sensor) and event-based recordings (record is made when the moving entity performs an activity).

### 2.2.2 Scale and data uncertainty in movement data

Scale matters in movement mining (Laube 2014). As Montello (2001) emphasizes, analysis scale, the scale at which phenomena are measured or aggregated, should match phenomenon scale, the scale at which geographic phenomena occur. The word "aggregated" in the previous sentence does imply that once the data is collected, the analysis does not necessarily need to have the same scale. Laube & Purves (2011) showed, using cross-scale analysis, that derived movement parameters of a cow's movement, e.g. its speed, take different values depending on the analysis scale of the measurement. They thereby stress the importance of matching analysis to phenomenon scale even after completed data acquisition.

As every kind of data, also movement data is prone to uncertainties (Andrienko et al. 2008). The fact that in movement analysis the movements have to be sampled in some temporal and spatial interval, and that whatever happens in between is unknown or at least uncertain is one of the fundamental challenges of the field (Laube 2014). Another important factor is the positional accuracy of the movement fixes but also the accuracy of the context information (Imfeld et al. 2006). The influence of

these positional accuracies on the results of a movement mining task furthermore depends on its nature (Imfeld et al. 2006). A point-in-polygon test for example reacts more sensitively to positional inaccuracies than aggregation tasks (Imfeld et al. 2006).

### 2.2.3 Movement mining tasks

Movement mining tasks can be grouped into four classes: Segmentation and filtering, similarity and clustering, movement patterns, and exploratory analysis and visualization (Laube 2014).

Segmentation of a trajectory entails dividing it into a number of subsequences (Laube 2014). These subsequences can then be analyzed further so the segmentation can be seen as a filtering/pre-processing step (Laube 2014). As an example, Spaccapietra et al. (2008) divide raw movement data into stops and moves, which can subsequently be assigned to a specific activity.

The computation of trajectory similarity and subsequent clustering is non-trivial as in a first step it has to be decided what to compare (Laube 2014). Should whole trajectories or just segments of them be compared? Should trajectories be handled as simple geometries without temporal information or should they be analyzed as space-time traces? In a second step the most appropriate of many possible distance metrics, including Euclidian distance, Fréchet distance, longest common subsequence and optimal matching distance, needs to be chosen (Laube 2014). Finally the most appropriate of many available clustering algorithms, including k-means, self-organizing trees and agglomerative nesting, has to be selected (Laube 2014). Nanni & Pedreschi (2006) use a density-based clustering algorithm relying on Euclidian distance of trajectory-segments computed by a time focusing algorithm that allows just using informative and similar subsequences of the trajectories for the clustering. Buchin et al. (2011) use Euclidian distance to compare trajectories, but they do so the other way around as they conduct the similarity analysis first, to find the longest similar subsequences of two trajectories, in which the trajectories are subsequently partitioned (Figure 7). Du Mouza & Rigaux (2005) highlight the advantages of discretized movement spaces for computational trajectory comparison (Figure 8), and Kang et al. (2009) show how trajectory similarity in cellular space can be assessed using the two measures Longest Common Subsequence and Common Visit Time Interval (Figure 9). Pelekis et al. (2012) combine primitive and derived movement parameters to a trajectory similarity measure they call locality in-between polylines and then use visual analytics tools to make sense of these measures (Figure 10). Dodge et al. (2012) also compute derived movement parameters (such as speed, acceleration, direction) but then use a similarity measure called Normalized Weighted Edit Distance, which basically compares trajectories as abstract sequences (Figure 11). This abstract sequence comparison cannot only be done with derived movement parameters, but also with *raw* trajectories, which is portrayed in Section 2.3 of this thesis.

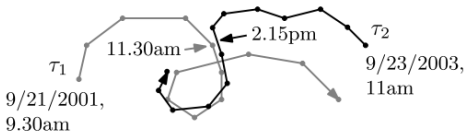



*Figure 7: Longest similar subsequence in the middle of T1 and at the end of T2. Source:* Buchin et al. (2011).

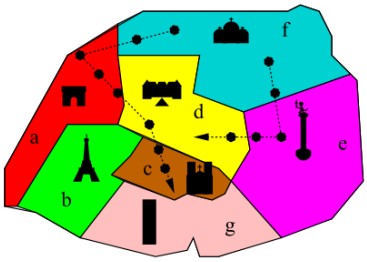



*Figure 8: Movements in a discretized reference space. Source:* du Mouza & Rigaux (2005).

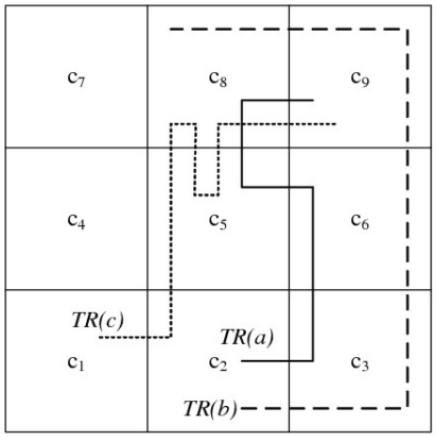



*Figure 9: Trajectories in a cellular reference space. Source:* Kang et al. (2009).

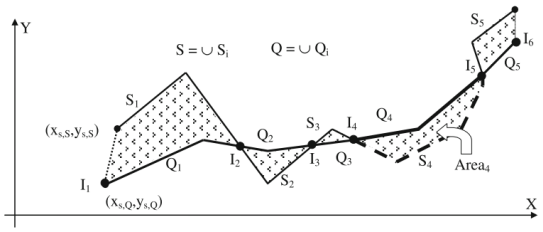
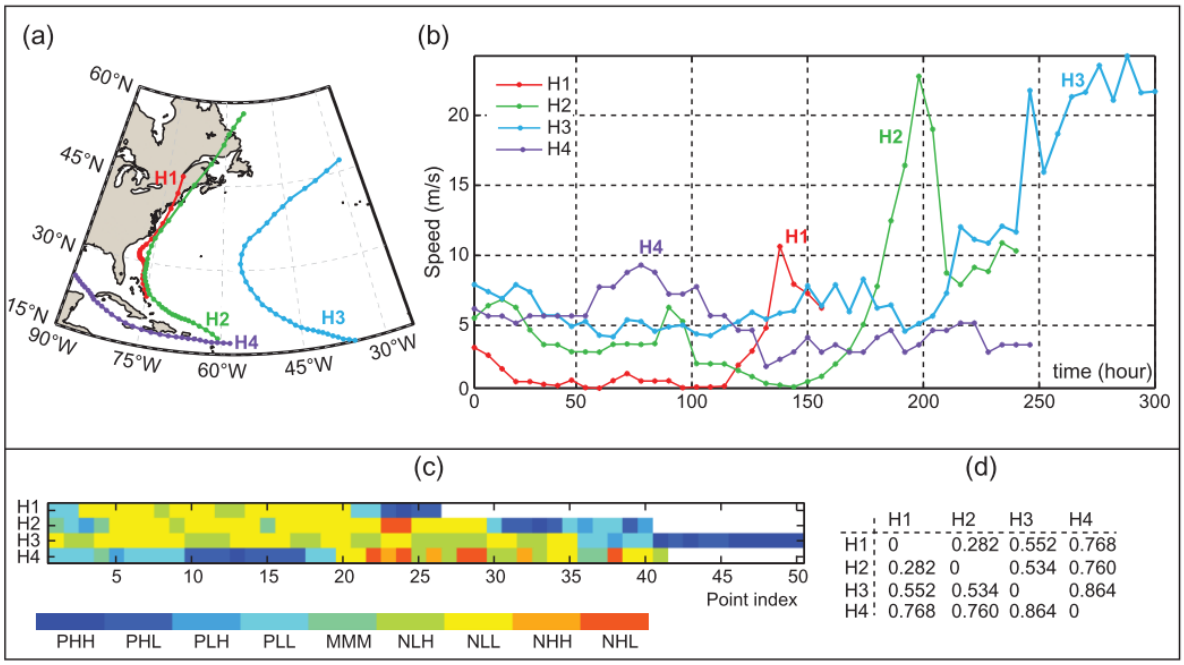



*Figure 10: Area between the lines (locality in-between polylines) as a trajectory similarity measure. Source:* Pelekis et al. (2012).




| | H1 | H2 | H3 | H4 |
|---|---|---|---|---|
| H1 | 0 | 0.282 | 0.552 | 0.768 |
| H2 | 0.282 | 0 | 0.534 | 0.760 |
| H3 | 0.552 | 0.534 | 0 | 0.864 |
| H4 | 0.768 | 0.760 | 0.864 | 0 |

*Figure 11: Four hurricane trajectories (a), their speed profiles (b), after segmentation (c) and trajectory distance computed with the Normalized Weighted Edit Distance algorithm. Source:* Dodge et al. (2012).

Following Laube (2014), the identification of understandable movement patterns is the quintessence of movement mining. Possible patterns include *flock leadership* or *moving clusters* (Laube 2014), but

also individual patterns such as recurrent movements. Movement patterns were therefore categorized into individual movement behaviors and dynamic collective behaviors by Andrienko & Andrienko (2007), whereas Dodge et al. (2008) focus on the complexity of the movement patterns and differentiate between primitive and compound patterns.

Mining for movement patterns can also happen in a non-deterministic way, for which exploratory analysis and visual analytics play an important role (Laube 2014). Considering the specific characteristics of movement datasets, tools such as small multiples, aggregation or interactive interfaces can be gainfully applied (Laube 2014).

### 2.2.4 Evaluation of movement mining results

Inconsiderate translation of data mining results into knowledge can easily lead to unsatisfying outcomes because of the meaninglessness or invalidity of identified patterns (Fayyad et al. 1996). Laube (2014) therefore proposes the evaluation of the found patterns by verifying, validating and credibility-testing them (following the terminology of Rykiel 1996) and by assessing the efficiency of the algorithm used to discover the patterns.

Verification focuses on the technical implementation of analytical ideas (Laube 2014). How good is the initial idea represented in the assembled formulas or in the written computer code (Laube 2014)?

Validation reviews whether the results of a verified method lie satisfyingly close to what they should be (Laube 2014). According to Laube (2014), methods to achieve this include expert interviews, visualization techniques, sensitivity analysis for parameter settings and comparison of the results to results of established methods with the same aim.

The credibility of results describes whether potential users of them can believe in and possibly also apply the found patterns in some way (Laube 2014). Laube (2014) further links credibility to Silberschatz & Tuzhilins (1996) concept of interestingness. Following Silberschatz & Tuzhilin (1996), data mining results are interesting for a user if they are either unexpected or if the user can plan a future action because of them, called *actionability*.

Lastly, Laube (2014) stresses the importance of efficient movement mining algorithms, as the efficiency determines the scalability and general usability of an algorithm in a given research context.

If no ground truth is available the presented evaluation techniques are harder to apply, but evaluation is by no means impossible. Zafarani & Liu (2015) show how large social media datasets, accompanied by no or very little ground truth can be evaluated using methods such as causality detection (evaluating the why of things) and outcome evaluation (evaluating the how of things).

## 2.3 SEQUENCE ALIGNMENT METHODS

Sequence alignment methods are techniques to arrange multiple sequences to find similarities in them. Sequence alignment can be used as a movement mining tool but originally it was developed by microbiologists. Section 2.3.1. shortly discusses the origin of the method and the various non-spatial uses, after which Section 2.3.2 portrays a number of spatial studies performed with sequence alignment methods. The sequence alignment technique applied in this thesis is explained in detail in the methods section 4.1.

### 2.3.1 Origin and non-spatial uses

In the 1970s, microbiologists started to analyze sequences of amino acids and nucleotides to investigate the human proteins and DNA (Mount 2004). Multiple strings of nucleotide symbols were compared to find similarities and differences between two DNA sequences (Mount 2004). As sequential data accumulates not only in microbiology, the technique was introduced to other fields in fast succession: Abbott & Forrest (1986) introduced sequence alignment methods into the social sciences by investigating the historical sequences of dance traditions of English villages in the 18th and 19th century. Chan (1995) studied career mobility of Hong Kong residents by investigating their job sequences. Wilson (2001) worked on a finer temporal scale level when he researched daily activity patterns of Canadian women by aligning their everyday activity sequences. Sequence alignment in the field of psychology was performed by Poole & Holmes (1995) when they observed decision making sequences with and without computer assistance and in the field of finance and marketing. Prinzie & Van den Poel (2006) analyzed sequences of customer retention efforts of an international bank.

### 2.3.2 Spatial sequence alignment studies

Bargeman et al. (2002) were the first to bring some sense of location into SAM studies by categorizing places according to their functions. They studied vacation behavior of Dutch tourists as sequences of staying at home, domestic vacations and vacations abroad. Bargeman et al. (2002) conclude that this approach enables them to derive a typology of the tourist concerning the *"frequency, duration, timing, destination, temporal and spatial sequence, and spatial repetition of their travels"* (Bargeman et al. 2002, p. 333). Similarly, but with an even looser link to geographic locations, Stovel & Bolan (2004) analyzed residential mobility trajectories between rural, suburban and metropolitan areas which allows them to discuss push- and pull factors of a given place type for a given population segment (e.g. young families may like to live in the countryside, students prefer the city). Shoval & Isaacson (2007) went one step closer to geographic location, compared to the *functional places* discussed before, when they investigated tourist tracks through the medieval Israeli town of Akko. They divided the town area into polygons (Figure 12), each containing one tourist attraction and then analyzed sequences of these polygons from 139 visitors equipped with GPS-trackers. As a result of the sequence alignment, the visitors were grouped into three groups, each showing a distinct trajectory type, which is displayed in Figure 13 in the form of the typical space-time path of each group.


*Figure 12: Partitioning of Akko into polygons, each containing one touristic attraction. Source:* Shoval & Isaacson (2007)

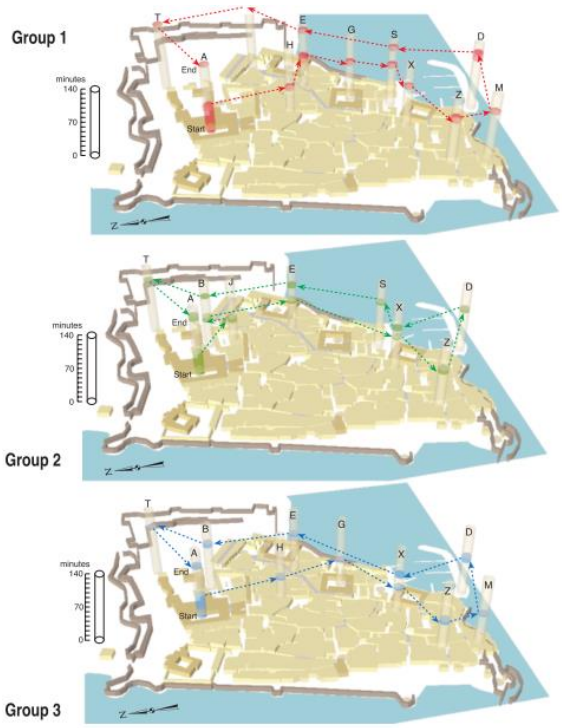



*Figure 13: Typical trajectories for the three types of visitors visualized on a 3D map. Source: Shoval & Isaacson (2007)*

In a later study, Shoval et al. (2015) aligned trajectories of tourists visiting Hong Kong with a similar methodology as in the Akko study. Out of the 261 participants, fifteen groups were computed based on a taxonomic tree of trajectory similarity (Figure 14) and typical sequences for each group were computed (Figure 15). Subsequently these groups were compared with demographic characteristics and background of the participants such as length of stay in Hong Kong, companions, age, gender and income.

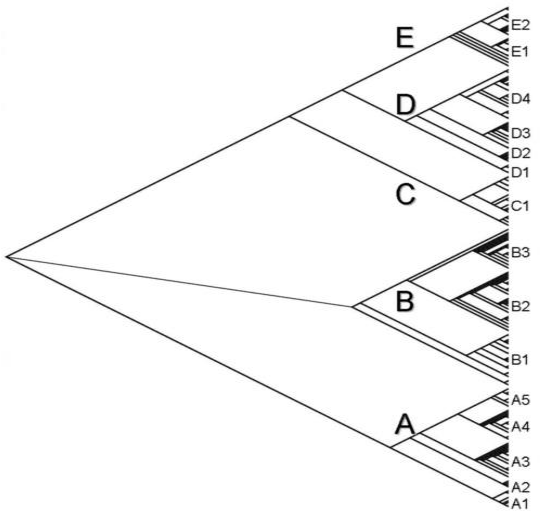



*Figure 14: Taxonomic tree of trajectory similarity of Hong Kong visitors. Source: Shoval et al. (2015)*

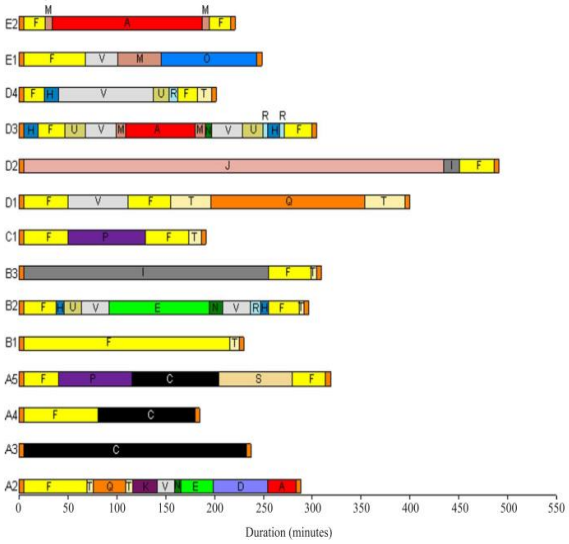



*Figure 15: Typical sequences of computed groups of Hong Kong tourists. Visualized with colors and letters standing for polygons laid over the town area. Source: Shoval et al. (2015)*

The relation between these demographic variables and the trajectory group membership was investigated using Chi-Square tests, resulting, partly because of the small sample size, in no significant relationships. Nevertheless, indicative inferences such as the findings that young age makes a visit of the Disney land more probable or that on the last day of their stay visitors tend to remain close to their hotel could be made.

The application of SAM in an indoor environment was introduced by Delafontaine et al. (2012) when they aligned trajectories of visitors of a big trade fair in Belgium obtained through Bluetooth tracking. 22 Bluetooth nodes with a radio range of 20 m were spread more or less regularly over the eight exhibition halls, which covered an area of about 56'000 $m^2$ (Figure 16)(Delafontaine et al. 2012). Over the course of the 5-day trading fair, movement sequences of 14'498 unique devices were recorded and subsequently aligned. The alignment resulted in 21 clusters identified by their median and average sequence, which minimizes the sum of Levenshtein distance (median) or the sum of squared Levenshtein distances (average) to all the sequences in the group (Figure 17).

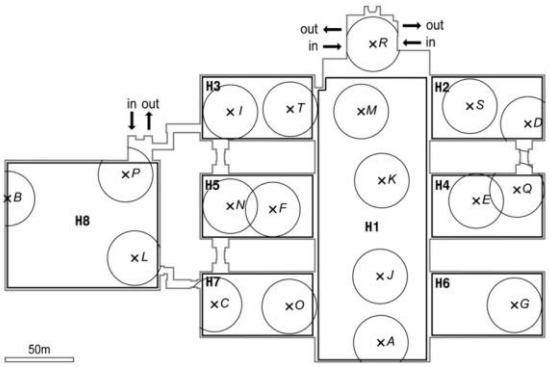



*Figure 16: Floor layout with hall names (H1-H8) and Bluetooth nodes (A-T). Source: Delafontaine et al. (2012)*

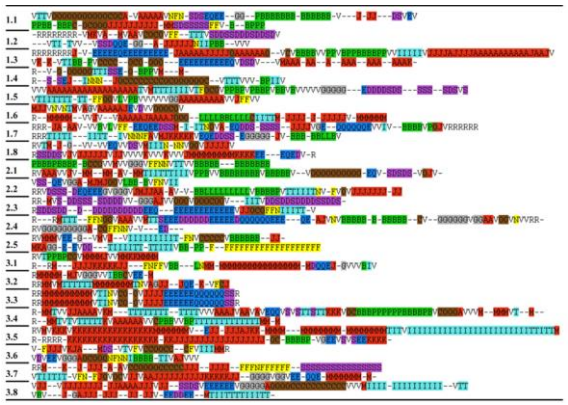


*Figure 17: Median (1st line) and average (2nd line) sequence per trajectory cluster color coded according to halls. Source: Delafontaine et al. (2012)*

Delafontaine et al. (2012) then analyzed these sequences qualitatively and found for example that most of the groups only visit each hall, except hall 1, only once, making most of their visit to one hall and not returning to it later.

Wilson (2008) went one step further in modeling space in SAM studies by using it directly and not as functional spaces or region polygons. Activities at a given location (geographic coordinate) are aligned according to the similarity of the activities and the similarity of the locations, which is modeled as the Euclidian distance in between the two locations. Yuan & Raubal (2014) also use Euclidian distance as a cost parameter when they analyze mobile phone user movements of 844'784 participants obtained from call detail records. They furthermore add absolute time as a dimension to SAM. Time was either used as an additional cost parameter or as a constraint for trajectory partitioning. Lee & Joh (2010) showed that SAM is not limited to the dimensions of time and space when they analyzed tourism behavior in Seoul and aligned activity, place, travel mode, accompanying status and absolute time. With this approach they were able to detect sequence patterns such as visiting the palace in old downtown in the morning and then visiting the palace in old downtown in the afternoon, and then shopping in Namsan/Yongsan (Lee & Joh 2010). That SAM cannot only be used for analyzing human movement but also to study animal movement, e.g. the study of deer habitat sequences, was shown by De Groeve et al. (2015). Furthermore, Çöltekin et al. (2010) showed that when using SAM space does not need to be modeled strictly geographically by analyzing gaze path sequences of participants interacting with a digital map as did Jiang et al. (2012) by studying web usage patterns across sites in the web space.

The output of all the presented papers are clusters of similar trajectories or trajectory similarity measures obtained from SAM, but all the researchers struggle to quantitatively evaluate the correctness of their results. They either call this problem future work or refer to Wilson (2006) who, using Monte Carlo simulations, at least showed that results of SAM perform better than random but still thought that SAM is very dependent on careful implementation and interpretation by the researcher.

## 2.4 IDENTIFICATION OF RESEARCH GAPS

Sequence alignment can be classified as a relatively new and only partially well-researched movement mining approach. In the last ten years SAM was used indoors (Delafontaine et al. 2012) and outdoors (Shoval & Isaacson 2007; Shoval et al. 2015; Lee & Joh 2010), analyzing movements of tourists (Shoval

& Isaacson 2007; Shoval et al. 2015; Lee & Joh 2010), sequences of everyday activities and movements (Wilson 2008; Yuan & Raubal 2014), animal movement (De Groeve et al. 2015), gaze paths (Çöltekin et al. 2010) and web usage (Jiang et al. 2012). Large (Yuan & Raubal 2014; Jiang et al. 2012) and small (Shoval & Isaacson 2007; Delafontaine et al. 2012) datasets originating from GPS tracking (Shoval & Isaacson 2007; Shoval et al. 2015; De Groeve et al. 2015), Bluetooth tracking (Delafontaine et al. 2012), mobile network tracking (Yuan & Raubal 2014), eye tracking (Çöltekin et al. 2010), and questionnaires (Wilson 2008; Lee & Joh 2010) have been analyzed.

SAM has, to my knowledge, not yet been applied to data collected through Wi-Fi tracking and neither to movement data in a shopping mall environment. The study of Delafontaine et al. (2012) analyzing a Bluetooth tracking dataset of a fair comes closest to the scope of this thesis, nevertheless, differences originating from the data collection (smaller range of Bluetooth compared to Wi-Fi, less people having Bluetooth on, leading to a smaller population) and the different environments (different level of space restraints in fairs compared to shopping malls and different kinds of indoor movement behavior) exist. An additional contribution this thesis tries to make addresses the gap identified by Shoval et al. (2015) when they state that *" …[an] area of importance yet to be explored is the impact of the spatial and temporal scale on the outcome of the alignment"* (Shoval et al. 2015, p. 91). All reviewed SAM studies accepted the data-inherent scale as the phenomenon scale and thereby optimal analysis scale. This thesis tries to investigate this specific research gap in depth.

# 3 DATA AND PRE-PROCESSING

All the analysis conducted throughout this thesis is based on Wi-Fi logs from the Westfield shopping center in Sydney, Australia. Why and how this dataset was received and what has already been done with it is described in Section 3.1, details and statistics of the data are presented in Section 3.2, limits of the dataset are listed in Section 3.3, how the data was pre-processed is explained in Section 3.4 and the categorization of the trajectories based on their temporal characteristics is described in Section 3.5.

## 3.1 RECORDING AND HISTORY OF THE DATASET

In 2013, the TRIIBE project (TRacking Indoor Information Behaviour)[1] received access to the comprehensive Wi-Fi logs of the Westfield shopping center in Sydney, Australia, to study the web/physical behavior of indoor shoppers. Figure 18 shows a 3-D model of the Westfield shopping center. It spans 90'000 $m^2$ over seven levels (level 1-4 featuring mainly shops, the food court on level 5 and mixed use on the smaller levels 6 and 7) is home to over 200 shops and is equipped with 69 Wi-Fi APs. The mall-Wi-Fi is free for registered users, which by registering give their consent for the collection of their data.

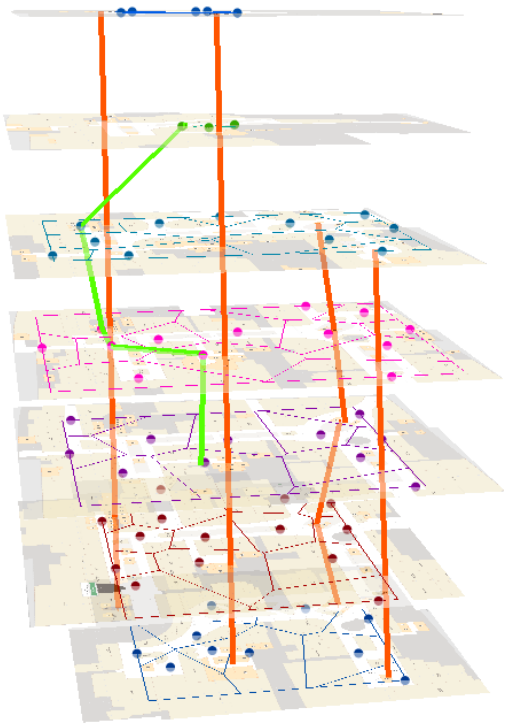


*Figure 18: 3-D Model of the seven levels of the shopping center with APs (points), Voronoi regions (polygons), lifts and escalators (orange lines) and an example trajectory (green line).*

The data was collected between September 2012 and October 2013. The collection was conducted in-network, entailing all the characteristics of such tracking described in Section 2.1.1.1. The dataset consists of three different logs: The Wi-Fi access point log, the browsing log and the query log. The access point log consists, among other for this thesis less important information, of the username of the registered user, the MAC-address of the connected device, the name of the AP the device is

[1] http://tomko.org/research-projects/

connected to and the time of association. The browsing log stores the online behavior of the users and the query log is an excerpt of the browsing log, concentrating on query behavior. The two latter logs are not described in more detail, as they are not used in this thesis (for more detail see Ren et al. (2015)).

The dataset has already been used in multiple studies by the TRIIBE team. Ren et al. (2014) used it to investigate the impact of spatial context on the information behavior of shoppers. They categorized the shops into different groups (e.g. Men's fashion, jewelry, groceries…) and found that information behavior in locations belonging to different shop categories was significantly different. On the other hand, information behavior at locations in the same category was shown to be similar. Ren et al. (2015) observed the web behavior in indoor retail space in more detail and found, amongst other findings, that people who visit the mall repeatedly tend to visit similar places and similar web content during their repeated visits and that accompanying users tend to show similar web behavior. Tomko et al. (2014) made initial findings about the spatial and temporal patterns of users, but they are limited to temporal return patterns to the shopping center over the course of a year, which is why trajectories of spatio-temporal movement behavior inside the walls of the building are up to now unresearched.

## 3.2 CHARACTERISTICS OF THE TRACKING DATASET

The Wi-Fi association log consists of 907'093 fixations with APS. Each fixation features extensive information about the device and its communication with the network, of which the attributes used for this study are shown in Table 1 and described in the following:

- CLIENT_USERNAME: Username the visitor uses to log on to the network. It can therefore be used as unique, persistent ID of the user enabling to track the user across sessions, and across devices. To ensure user privacy, this attribute was non-reversibly hashed.
- TRAJECTORY_ID: A derived attribute consisting of a unique combination of CLIENT_MAC, CLIENT_USERNAME and the day part of ASSOCIATION_TIME. In this sense, a trajectory incorporates all fixations of a user with one device over the course of one day.
- CLIENT_MAC: Unique ID of the network card in the internet enabled device, which can be used as ID of the device in the study context. The MAC-address is non-reversibly hashed to preclude conclusions about individuals which would violate data privacy principles.
- ASSOCIATION_TIME: Timestamp of the communication with the AP, used to create sequences of positions, known as movement.
- AP_NAME: The ID of the AP the device is connected to. This attribute is used for positioning in this thesis.

| CLIENT_USERNAME | TRAJECTORY_ID | CLIENT_MAC | ASSCOCIATION _TIME | AP_NAME |
|---|---|---|---|---|
| @M5IMFJPVVCI7EBAFOQ179TE72A@ | $13J5E4PMT01793MU74HI63GCFI$ @M5IMFJPVVCI7EBAFOQ179TE72A@2013-08-13 | $13J5E4PMT01793MU74HI63GCFI$ | 13.08.2013 16:15 | wau-nswsyd-wap021 |
| @LK8PDGQER89E9ARO598LIO01JI@ | $RURPL35SM3QI1UMGQC5SRSUGEF$ @LK8PDGQER89E9ARO598LIO01JI@2013-08-14 | $RURPL35SM3QI1UMGQC5SRSUGEF$ | 14.08.2013 12:35 | wau-nswsyd-wap-skybeam |
| @QJGIFTEJ30F79REAQR402LHOIM@ | $K4I1IV1VNR5592BV4504KC5S0T$ @QJGIFTEJ30F79REAQR402LHOIM@2013-08-13 | $K4I1IV1VNR5592BV4504KC5S0T$ | 13.08.2013 17:10 | wau-nswsyd-wap021 |
| @52BOF10PH5B0VL7QKTHV44APC5@ | $OJ5V0VS2JVL42C8LNG5V64VI57$ @52BOF10PH5B0VL7QKTHV44APC5@2013-08-13 | $OJ5V0VS2JVL42C8LNG5V64VI57$ | 13.08.2013 17:05 | wau-nswsyd-wap033 |
| @S7KDOHKN88BQBROLRGN5POQCE@ | $U1I132C5RRIBHF1T7JPK4VCOP8$ @S7KDOHKN88BQBROLRGN5POQCE@2013-07-11 | $U1I132C5RRIBHF1T7JPK4VCOP8$ | 11.07.2013 22:11 | wau-nswsyd-wap049 |
| @S7KDOHKN88BQBROLRGN5POQCE@ | $U1I132C5RRIBHF1T7JPK4VCOP8$ @S7KDOHKN88BQBROLRGN5POQCE@2013-07-11 | $U1I132C5RRIBHF1T7JPK4VCOP8$ | 11.07.2013 19:25 | wau-nswsyd-wap014 |

*Table 1: Except from the data with the attributes important for this thesis*

The 907'093 fixations can be aggregated to 260'296 trajectories, which are in average between 3 to 4 fixations long. The 260'296 trajectories come from 120'548 unique users, which results in an average of just over 2 visits per user, which however is by no means normally distributed (see Figure 24).

Ren et al. (2015) described the dataset in more detail concerning temporal, spatial, social and web-activity characteristics. The key figures most important for this thesis are recapitulated here:

- **Temporal patterns of users' visits**: At the scope of one day, it was found that visitors' presence correlates strongly with the opening hours of the shopping mall. At the scope of one week, Thursday, the traditional shopping day in Australia, was found to attract most visitors to the mall (17%, compared to 12%-15% the other days). (Ren et al. 2015)
- **Length of visits to the mall**: The time between the first and last association with the web during one day showed an irregular distribution, with the majority (66%) of visits being between 3-4 hours long (Figure 19). (Ren et al. 2015)
- **Frequencies of repeat visits:** 67% of the visitors were only registered once in the mall Wi-Fi. Many of the other visitors who came to the mall repeatedly showed a multiple of 7-day interval between visits (Figure 20). (Ren et al. 2015)
- **Spatial context:** Ren et al. (2015) categorized the mall into three spatial contexts: The food-court context (containing 11 APs), the retail context (46 APs) and the navigational context (10 APs). Table 2 shows that most of the association time is spent in the retail context, followed by the food-court with a high percentage per AP, leaving the navigational context with the clearly smallest part of the association time.

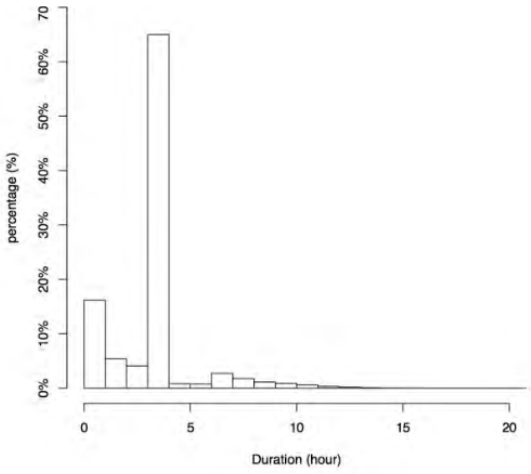



*Figure 19: Length of visits to the mall. Source: Ren et al. (2015).*

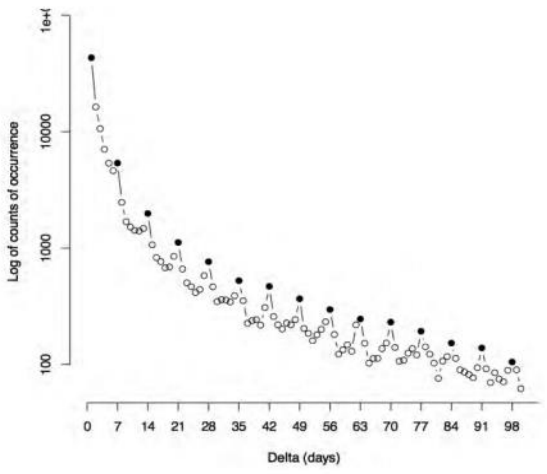



*Figure 20: Visitor return pattern. Source: Ren et al. (2015).*

| Context | % of assoc. time | Avg. time per visit [h] |
|---|---|---|
| Food-court | 23% (2.06% per AP) | 1.39 |
| Retail | 70% (1.52% per AP) | 2.29 |
| Navigational | 7% (0.68% per AP) | 1.00 |
| Total | 100% (1.49% per AP) | 2.77 |

*Table 2: Spatial context of associations. Source: Ren et al. (2015).*

Despite mobile tracking data being one of the most used examples of Eulerian movement data (Laube 2014), this category does not fit the dataset perfectly, as movement is not recorded when the moving

entity crosses borders or gates but in a (irregular) time interval. In fact, the movement data category "event-based records" proposed by Andrienko et al. (2008) fits best to the dataset, as a user's position is recorded whenever he/she interacts with the mall Wi-Fi. Regarding movement space, the mall visitors in the dataset move through a highly constrained environment. The floor layout and especially the relatively few level transitions have a strong channeling effect, which has to be taken into account when analyzing the movements. Furthermore, the discrete nature of the dataset as well has an influence on the selection of appropriate analysis tools, as algorithms working well for continuous movement datasets not necessarily produce the desired results applied to discrete movement datasets.

## 3.3 LIMITS OF THE DATASET

The Wi-Fi tracking system can only track people carrying devices that are actually connected to the shopping mall's Wi-Fi. Furthermore, the network only covers common places such as halls and hallways and not for example the inside of shops. People who are connected to different WLANs, use their own cellular data or do not carry a web enabled device at all cannot be tracked. Moreover, the trajectories collected do not need to capture whole shopping trips, as visitors may turn Wi-Fi on/off during their visits. For people that can be tracked two main limitations persist: The spatial resolution and the gaps in the trajectories.

### 3.3.1 Spatial resolution of the tracking dataset

The location of the APs and their associated adjusted Voronoi polygons serve as a proxy for the location of the users. Movement can therefore only be detected between these polygons and not within them. Furthermore, in some cases two slightly different paths between two neighboring polygons are possible and with our dataset it is impossible to find out which path has been used (Figure 21).

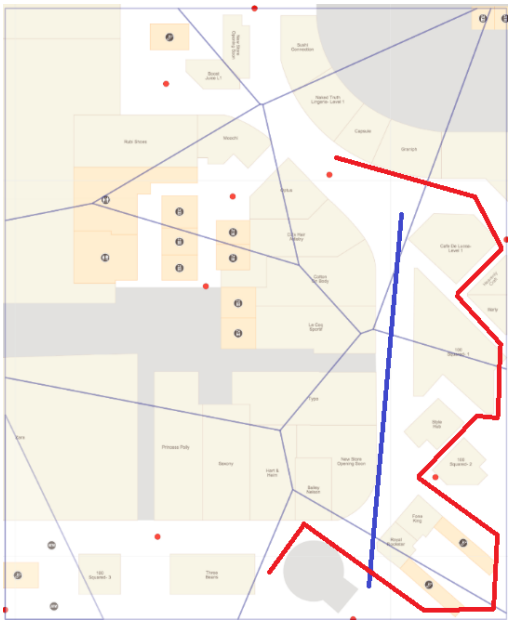


*Figure 21: Two different paths showing the same sequence of visited AP-regions.*

### 3.3.2 Gaps in the trajectories

Mobile phones fall into sleep mode when they are not used for a couple of seconds, and when they do so, they can no longer be seen by Wi-Fi tracking system. Several applications that send regular pings

to the network can prevent this from happening, but nevertheless many of the recorded trajectories have gaps of up to several hours. Supporting this observation, a large number of recorded transitions occur between spatially disjoint regions/APs. As it is not physically possible to accomplish such transitions between two spatially disjoint regions, this is a limit to the data that has to be considered. This limitation can be interpreted as a consequence of studying real life data, which, in contrast to more instrumented studies, often comes with less complete trajectories.

Another limitation of the data is, that the highest possible temporal sampling rate of the AP-log is only 5 minutes, meaning that in the whole dataset no two fixations from the same user within a period of five minutes exist. Whether this limitation comes from the data recording phase or from the post-processing by the data provider could not be established. The average gap length, meaning the time between two AP fixes, over the whole dataset is 29.1 minutes.

## 3.4 PRE-PROCESSING

As the data was already processed and cleaned up to some degree by the TRIIBE team, the main goal of my data preprocessing was to find an applicable subset of the data for further use with the sequence alignment methods, to test my research questions. This subset should be of appropriate size, large enough to still be able to detect statistical relations in it and small enough to enable practical computation times, and the entries should be chosen in a way that the influence of the before mentioned limitations of the data can be minimized. The data was delivered as a MYSQL database and pre-processed in R.

In the following two filters are applied to the data and the effects of these filters on the remaining subsets are presented.

### 3.4.1 Minimum trajectory length filter

For successful further analysis trajectories as long as possible are beneficial. Figure 22 shows the histogram of trajectory lengths (number of AP-fixes). As a trade-off between subset size and trajectory length (trajectories should be long enough to carry information about the movement), the cut off level of 5 was chosen, which reduces the size of the datasets as shown in **Error! Reference source not found.**.

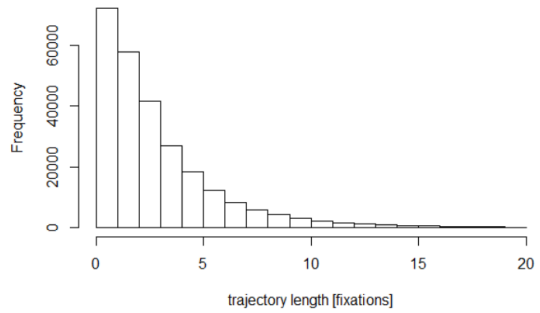



*Figure 22: Histogram of trajectory length*

| | Trajectories | Unique users |
|---|---|---|
| **Whole dataset** | 260’296 | 120’548 |
| **Trajectory length >=5** | 88’878 | 49’700 |

*Table 3: Trajectory length subsets*

### 3.4.2 Maximum average gap length filter – temporal resolution

To enhance movement mining results, the trajectories should include as few and as short gaps as possible and therefore a high temporal sampling rate is preferable. Figure 23 shows the histogram of average gap length (time between AP-fixes).

As a trade-off between subset size and trajectory length the three different cut off levels 5, 7.5 and 10 minutes were chosen, which reduces the subset as shown in Table 4. An average gap length of 7.5 in a trajectory of length 6, for example, could stand for three 10 minute gaps and three five minute gaps, but also for five 5 minute gaps and one 20 minute gap. Trajectories with only one fixation were filtered out as well.

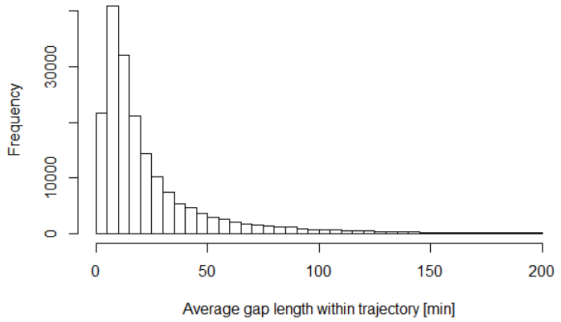



*Figure 23: Histogram of average gap length within trajectory*

| | trajectories | Unique users |
|---|---|---|
| **Whole dataset** | 260’296 | 120’548 |
| **Gap length <=10 min** | 62’674 | 42’904 |
| **Gap length <=7.5 min** | 43’171 | 31’170 |
| **Gap length <=5 min** | 21’730 | 16’701 |

*Table 4: Average gap length subsets*

These three subsets regarding the average gap length are used in the thesis to assess the influence of temporal resolution of the data on the movement mining results.

### 3.4.3 Combination of the filters

The final subsets that were used in the following analysis were computed through the combination of the two filters. The respective subset sizes are shown in Table 5.

| | Filters | Trajectories | Unique users |
|---|---|---|---|
| **Subset 1** | Trajectory length >=5 min, Gap length <=10 min | 15’241 | 12’785 |
| **Subset 2** | Trajectory length >=5 min, Gap length <=7.5 min | 7’223 | 6’466 |
| **Subset 3** | Trajectory length >=5 min, Gap length <=5 min | 282 | 274 |

*Table 5: Final subsets computed by combining the trajectory length and gap length filter.*

As users coming to the shopping center more than once are important for later analysis steps, Figure 24 shows the frequencies of repeated visits in Subset 2.

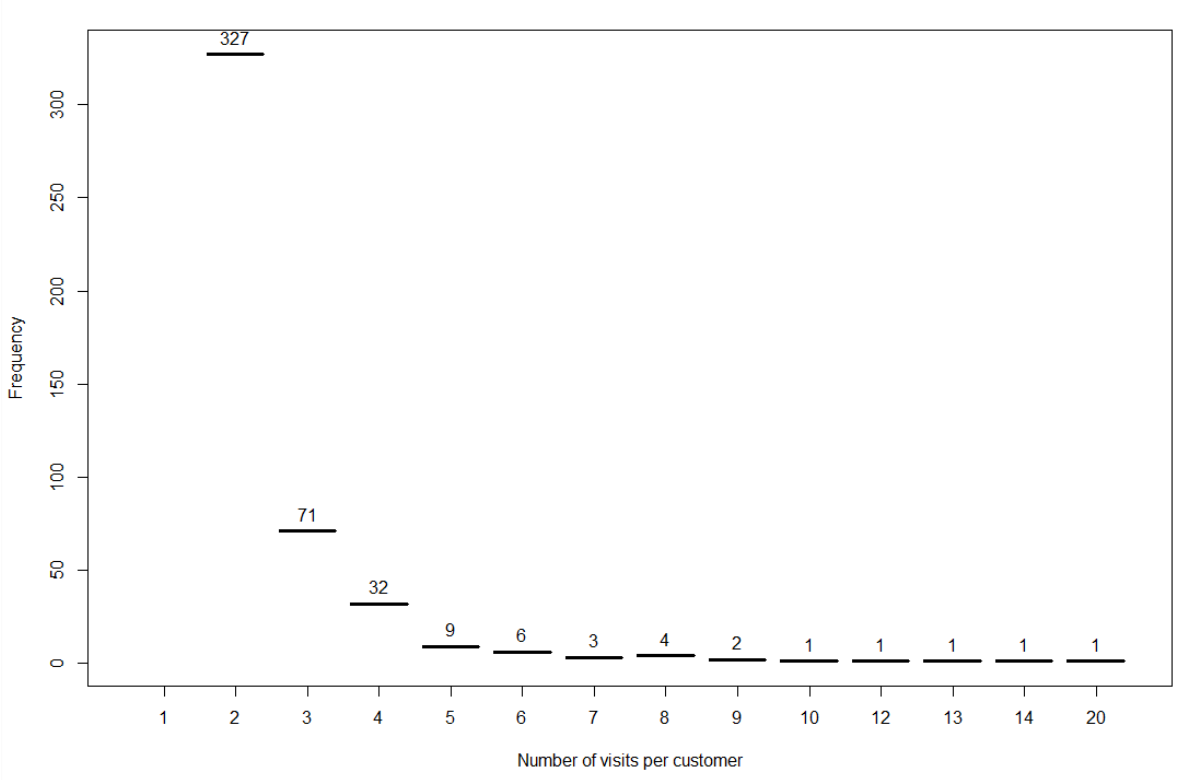
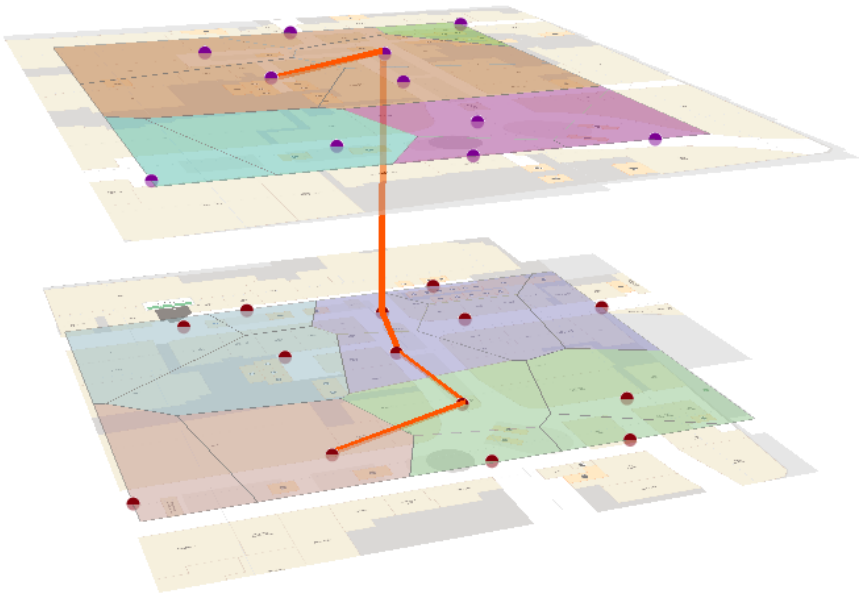



*Figure 24: Repeated visits in subset 2 (without visitors who came only once)*

### 3.4.4 Spatial resolution

In the last step of the pre-processing, the three computed subsets were geo-coded using three different scale levels: APs, areas and levels.

The highest achievable spatial resolution is the data-given AP service region. Movement is modeled as a sequence of all of the 69 hotspots the shoppers visited. At the area-level, the 69 hotspots were aggregated to 23 self-contained areas, such as wings of a building, that allow to decompose the area of a floorplan at a single floor level into a number of cohesive areas. Movement was then modeled as a sequence of areas. For the granularity of levels, the hotspots were aggregated to the 7 levels. Figure 25 shows how an example trajectory is expressed as a location sequence at the three presented scale levels.



Hotspots: h27-h32-h33-h38-h21-h25

Areas: a5-a6-a7-a7-a11-a11

Levels: l2-l2-l2-l2-l3-l3

*Figure 25: Example trajectory translated to location sequences at three different granularities of hotspots, areas and levels.*

## 3.5 VISITOR CATEGORIZATION BASED ON TEMPORAL CHARACTERISTICS

To give meaning to the groups identified by the sequence clustering discussed in Section 4.1-4.3, the influence of different characteristics of the visitors on the cluster membership of their trajectory is of interest. In an early phase of this thesis, it was planned to use a visitor survey from the TRIIBE project to assign demographic information to each user and his/her trajectories, which could then be used to create labeled groups that could be compared with the latent groups found by the sequence clustering. Due to privacy issues, the linkage of these two data sources could not be achieved and the user categorization had to be conducted based only on information of the Wi-Fi association log. However, the log contains information about the return habits of the visitors, which allows the deterministic categorization of the visitors based on parameters that are not part of the physical trajectory. It is then of interest, whether the physical trajectories for certain visitor categories are highly typical. Table 6 shows the categorization variables and the values the respective variable may take for all the users that came to the mall more than once.

| **Categorization variable** | **Values** |
|---|---|
| Time of day | {night (22-6),morning (6-12), afternoon (12-18), evening (18-22)} |
| Mode weekday/weekend | {weekday (mon-fri), weekend (sat-sun)} |
| Mode weekday | {mon, tue, wed, thu, fri, sat, sun} |
| Return period | {<week, 1-2 weeks, 2-3 week, 3-4 weeks, >4weeks} |

*Table 6: Visitor category based on return habits.*

Visitors were categorized based on their most common shopping time, their most common shopping day, whether they usually come during the week or at weekends and based on their average return period. Bar charts of the distributions of the four categorizations are given in Figures 26-29.

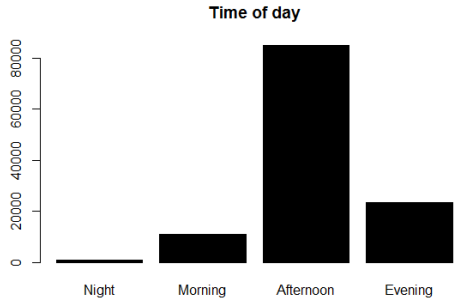



*Figure 26: Distribution of users categorized based on time of day.*

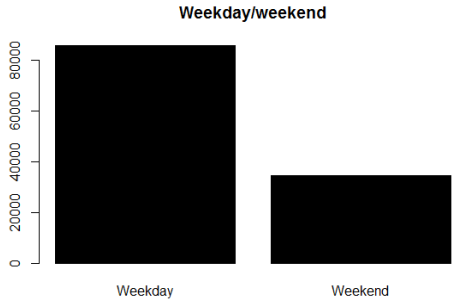



*Figure 27: Distribution of users categorized based on weekday/weekend.*

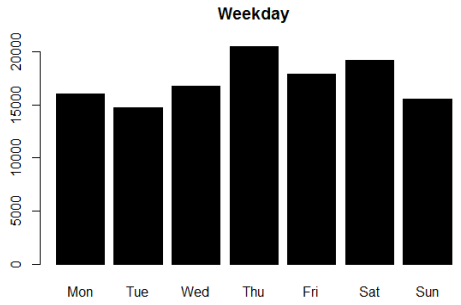



*Figure 28: Distribution of users categorized based on weekday.*

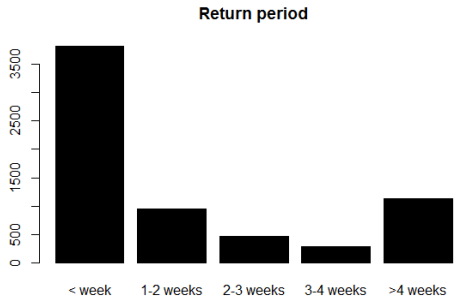



*Figure 29: Distribution of users categorized based on return period.*

The Mall Wi-Fi is used most during the afternoon (6h), followed by evening (4h) and morning (4h) and during the night (8h) the Wi-Fi is very seldom used. The visits per day are approximately the same for weekdays and weekends, with Thursday, the national shopping day in Australia, and Saturday being the busiest days. Most users that come more than once have a return period shorter than one week.

# 4 METHODS

Methods used in the course of this thesis include sequence alignment methods (Section 4.1), clustering methods (Section 4.2) and validation/calibration workflows (Section 4.3) and were applied in this order.

## 4.1 SEQUENCE ALIGNMENT

This chapter presents the two sequence alignment tools that were most used in the previously discussed literature, ClustalG and TraMineR. First, the main algorithm of the method, the optimal matching algorithm, is presented in Section 4.1.1. Second, the history and usage of ClustalG is presented in Section 4.1.2, together with the advantages and drawbacks of this software. The sometimes substantial drawbacks of ClustalG led to the decision to use the R-package TraMineR for this thesis, which is portrayed in Section 4.1.3.

### 4.1.1 Optimal matching distance

The optimal matching distance was originally invented by (Levenshtein 1966) and introduced to the social context by (Abbott & Forrest 1986). Optimal matching computes *"minimal cost, in terms of insertions, deletions and substitutions, for transforming one sequence into another"* ((Gabadinho & Ritschard 2009, p. 97). Insertion and deletion are often summarized into *indel*, as the insertion of a letter in sequence A has the same effect as the deletion of the same letter in sequence B. Following are a few examples of the optimal matching distance with an indel cost of 1 and a substitution cost of 1.

1. Hello
2. Hallo

"a" in sequence 2 is substituted with "e" at the substitution cost of 1, resulting in an optimal matching distance of 1.

1. Helloy
2. Hxello

"x" in sequence 2 is deleted and "y" is inserted at the indel cost of 1 each, resulting in an optimal matching distance of 2.

1. Hello
2. Hxellu

"x" in sequence 2 is deleted at the indel cost of 1 and "u" in sequence 2 is substituted with "o" at the substitution cost of 1, resulting in an optimal matching distance of 2.

The optimal matching distances computed and the subsequently retrieved patterns and clusters are strongly dependent on the calibration of the cost parameters for the indel and substitution operations (Lesnard 2010). If in example 1 the indel cost was 0.1, "e" in sequence two would be deleted and "a" inserted, resulting in an optimal matching distance of 0.2, which is cheaper than 1 for the substitution. If on the other hand example 2 would be calculated with indel cost 2, "x" in sequence 2 would be

substituted with “e”, “e” with “i” and “o” with “y”, resulting in an optimal matching distance of 3, which is cheaper than 4 for the two indels.

With the determination of the costs, it can be controlled whether more indels (preserving the events and warping the time) or substitutions (changing the events and preserving the time) should take place. Figure 30 shows the spectrum of the ratio between substitution cost and indel cost with the two extremes Hamming distance, for which only substitutions are allowed and longest common subsequence, for which only indels are allowed.

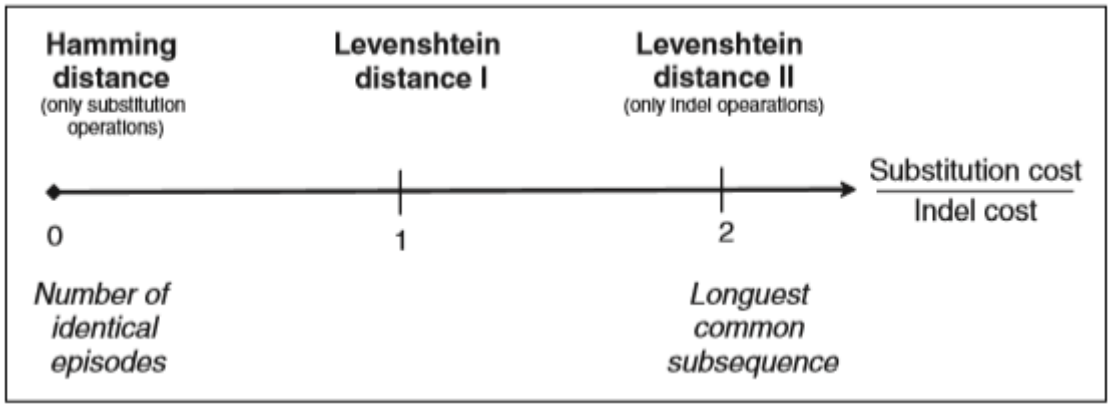



*Figure 30: Spectrum of the ratio between substitution cost and indel cost (Source: Lesnard (2010))*

The operation costs do not have to be constant and it is scientific consensus that substitution costs based on transition rates are most often more meaningful (Lesnard 2010). If the transition from state A to state B and in the other direction happens often, the distance between these two states is assumed to be small and the substitution of these two states should therefore not cost as much as a substitution of two more distant states with less transitions in between. The substitution cost based on transitions is calculated in TraMineR as follows:

$$2 - p(i|j) - p(j|i)$$

*Equation 1: Substitution cost dependent on transition rates (Source:* Gabadinho & Ritschard (2009)*)*

Where $p(i|j)$ is the transition rate from state $i$ to state $j$.

The appropriate value for the indel cost, or rather the appropriate ratio between indel cost and substitution cost is harder to establish. Lesnard (2010) proposes to not use indel operations at all, use Hamming distance with a transition rate substitution matrix, with the possible extension of including the time component in the substitution cost matrix, so that a substitution is not only dependent on the transition rates between the states to substitute but also at which point in the sequence this substitution happens. This is called the Dynamic Hamming Distance. Unfortunately the Hamming distances, dynamic and static, are not applicable with the visiting sequences used in this thesis, as only equal-length sequences can be compared. The calibration of the substitution cost/indel cost ratio thus has to be validated on the basis of the resulting distance matrices. As no obvious validation measures exist for distance matrices, they are validated on the basis of the goodness of the clusters that were

computed with them. This validation based upon a subsequent analysis step is following Lesnard (2010), who used entropy measures to validate different distance matrices for his research goal of identifying contemporaneous similarities.

### 4.1.2 ClustalG

#### *4.1.2.1 History*

The Clustal software family originates in bioinformatics where for example ClustalW (Thompson et al. 1994) was used to analyze nucleotide and peptide sequences. To this end, an alphabet size of 20 was totally sufficient as there only exist so many nucleic acids. This serious limitation of the software to spatial research was identified but put up with by Wilson (1998). One year later, Wilson et al. (1999) presented a rewrite of the original Clustal software called ClustalG. This version allowed an alphabet size of up to six letters to solve the problem of too few symbols for spatial research. Later, Wilson (2008) developed a version called ClustalTXY that even included Euclidian distances directly into the alignment and thereby facilitated multidimensional alignments.

#### *4.1.2.2 Usage*

ClustalG comes with a convenient graphical user interface that leads through the whole alignment process:

- Data Input: ClustalG reads sequences in seven formats, e.g. the Pearson (Fasta) format.
- Visualization and Editing: The sequences are displayed in the main window, which facilitates visual examination and even permits some rudimentary sequence editing.
- Alignment: Parameters such as type of alignment (local/global/approximate) and gap penalties (Gap Opening Penalty (GOP) and Gap Extending Penalty (GEP)) can be chosen and a similarity matrix can be loaded, in either a matrix or a parsing file format. The following alignment works without additional input and shows constant performance.
- Results: Four different types of results that are of use for this thesis can be generated. The first comes in the form of aligned sequences in the main window, which can be color-coded for better legibility. Second, an unrooted tree displaying similarity between sequences is produced as an output and can be viewed and analyzed using additional software such as e.g. NJplot. Additionally, log files of the pairwise and the multiple alignment can be saved and analyzed further.

#### *4.1.2.3 Advantages*

As a consequence of its graphical user interface, the usage of ClustalG is intuitive and only very limited programming skills (input pre-processing, post-processing for statistical analysis) are needed. Furthermore the resulting figures, the aligned sequences and the clustering tree summarize the results nicely and can directly be displayed and discussed in the thesis e.g. Delafontaine et al. (2012) and the log files provide an interface for statistical evaluation of the results.

#### *4.1.2.4 Drawbacks*

The user-friendly design of the software causes it to work like a black box in some cases: Boxes are ticked, rulers adjusted, buttons pressed and results come pouring out. The extensive help file explains most of these uncertainties regarding the functioning of the algorithms, but not all of them, as for example the clustering algorithm used subsequent to the sequence alignment remains a black box. The log files as interfaces for further statistical analysis (e.g. with R) represent another drawback of ClustalG. The size of the log files easily reachesexcessive dimensions (>10GB) such that a normal text

editor could handle the necessary pre-processing of the file so it can be read by R. The pre-processing therefore has to be done using command line “grep”-functions, which require pre-millennial programming skills and can be cumbersome.

In addition to these drawbacks, two bugs were identified in the early phases of working with ClustalG:

- When working with multi-letter alphabets, similarity matrices can only be loaded in the parsing file format. However, the reader of these parsing files was discovered to be erroneous in some cases, leading to wrong pairwise similarity scores. As a consequence of this, ClustalG could only be used with a one letter alphabet, which means a downgrade compared to the biological Clustal software, which again is not sufficient to capture the 71 hotspots used in this thesis.

- When applying local alignment, independently of the alphabet size, obviously wrong similarity scores result, presumably due to a wrong implementation of the Waterman-Smith algorithm. As a workaround, only the global alignment (Needleman-Wunsch algorithm) could be used, but nevertheless the reliability of ClustalG has to be questioned.

### 4.1.3 TraMineR

#### *4.1.3.1 History*

TraMineR is an R package supporting the visualization and mining of trajectories (Gabadinho & Ritschard 2009). The primary aim of the software is the mining of event or state sequences over the course of a life such as in work-school transitions or the sequence of marital statuses. However, the authors encourage the usage of the software with completely different sequential data, as for example movement data, as well.

#### *4.1.3.2 Usage*

TraMineR reads five different file formats, of which the state sequence (STS) is the most basic and the one used in TraMineR internally. In this format, the successive states of a sequence are given in consecutive columns, each column representing a fixed time unit (Gabadinho & Ritschard 2009).

The read sequences are then converted into *sequence objects*, which consist of the sequences equipped with several additional attributes like alphabet, color palette for the plots or code used for missing values.

The characteristics of these sequence objects can be described and visualized using the following plots and measurements (Gabadinho & Ritschard 2009):

- Sequence plots display all the sequences with custom color coding and ordered if requested. This and the other plots may also be computed for subsets.
- State distribution plots show the distribution of the states for each point in time (also available as table).
- Sequence frequency plots show the most frequent sequences, if specified with proportional bar width (also available as table).
- Transition rates can be computed, resulting in a matrix, in which each row represents the transition distribution of one given state into all the other states.
- Basic characteristics of individual sequences such as sequence length, state durations or number of transitions can be computed as well as composite measures of sequence

complexity such as turbulence (Elzinga & Liefbroer 2007), within-sequence entropy and complexity (Gabadinho et al. 2010) can be retrieved.

Of key interest for this thesis is the feature of TraMineR measuring similarities and distances between sequences. TraMineR supports basic metrics such as Longest Common Subsequence (LCS) but also the more complex optimal matching (OM) distance. The OM-distance is calculated as the sum of the indel-cost (constant and set by researcher between 0-2) and the substitution cost which is passed to the algorithm as a matrix. This matrix can be created manually in the case of expert knowledge or automatically taking the transition rates found in the data into account. The optimal matching in TraMineR uses the Needleman-Wunsch algorithm (Needleman & Wunsch 1970) and results in a distance matrix containing the pairwise similarities of all the sequences in the dataset. TraMineR offers capabilities to cluster the sequences based on these distances and to display dendrograms, however the number of provided clustering techniques is limited and cluster validation is not intended within TraMineR. Because of this, the clustering is made with a different R-package (clValid), using the distance matrix as an interface.

#### *4.1.3.3 Advantages*

TraMineR is, like most R package, very well documented, which ensures high transparency, and the customization possibilities regarding the core algorithms are numerous. The implementation only requires basic programming skills and the different results can each be plotted in an adequate way for later visual inspection and interpretation. As experienced in this thesis thesis, the usage of TraMineR is convenient as the pre-processing and the subsequent clustering and statistical analysis was also implemented in R, so no data interchange between different software products is needed.

#### *4.1.3.4 Drawbacks*

Most studies using TraMineR are social science studies indeed, but not from the spatial research field. With the exception of De Groeve et al. (2015) and Bleisch et al. (2013) no spatial studies using TraMineR were found, which demanded some degree of pioneer work during the implementation and for the interpretation of the results.

## 4.2 SEQUENCE CLUSTERING

„*Clustering is an unsupervised technique used to group together objects which are "close" to one another in a multidimensional feature space, usually for the purpose of uncovering some inherent structure which the data possesses"* (Brock et al, 2008, p. 1). Clustering can be used for data reduction, prediction based on groups, hypothesis generation and hypothesis testing, applications that are of importance in the fields of business, biology, web mining and spatial data analysis (Halkidi et al. 2001).

The one spatial application of clustering methods performed in this thesis is the clustering of movement traces, modeled as sequences. Once distances between sequences capturing their similarity are computed, the trajectories can be grouped so that the most similar ones are in the same group using a set of algorithms. These groups of similar trajectories then hint to latent groups of users.

Section 4.2.1 presents a non-exhaustive list of different clustering methods, and Section 4.2.2 discusses internal and external cluster validation measures.

### 4.2.1 Clustering methods

A multitude of clustering algorithms exists, serving the specific requirements of the different applications presented before. Jain et al. (1999) propose a categorization of the clustering algorithms into partitional clustering (directly decomposing the dataset), hierarchical clustering (successively aggregating data points (agglomerative algorithms) or dividing the dataset (divisive algorithms)), density-based clustering (clustering dense neighborhoods) and grid-based clustering (clustering on the basis of a space discretized into cells). In the following, a non-exhaustive list of clustering algorithms is provided. The clustering performance with the sequence dataset is later evaluated for all of these algorithms.

- **Model:**
  Model-based clustering is the most recent of the presented clustering algorithms and found high acceptance in the field of microbiology. It relies on Gaussian mixture models in which group memberships are computed using maximum likelihood. (Yeung et al. 2001)

- **SOM: self-organizing maps**
  SOM is an unsupervised, modular neural-network based clustering algorithm originating from the field of neurobiology. (Kohonen et al. 1997)

- **SOTA: self-organizing tree algorithm**
  SOTA, like SOM, is an unsupervised clustering technique relying on neural networks, but belongs to the class of divisive methods, in which one cluster containing all the observation is successively divided into finer clusters. (Dopazo & Carazo 1997)

- **PAM: Partitioning around medoids**
  PAM is a variant of the k-means algorithm, in which a fixed number of groups (*k* groups) is formed in the beginning and then data point switch groups until the lowest within-cluster distances are reached (compared to lowest within-cluster variances in k-means). (Kaufman & Rousseeuw 2009)

- **CLARA: Clustering Large Applications**
  CLARA samples the whole dataset into a number of sub-datasets and runs PAM on them. This results in faster computation times and hence allows clustering larger datasets. (Kaufman & Rousseeuw 2009)

- **FANNY: Fuzzy Analysis**
  FANNY applies fuzzy membership clustering in which each data point is allowed to have partial membership in each cluster. To reach a hard clustering result, each observation is assigned to the cluster in which it has highest partial membership. (Kaufman & Rousseeuw 2009)

- **AGNES: Agglomerative Nesting**
  AGNES, being a UPMGA method (Unweighted Pair Group Method with Arithmetic Mean), is, because of its intuitive understandability and user-friendly graphical interface (dendrogram), one of the most popular clustering techniques. Each data point is initially situated in its own cluster and is, based on a distance-matrix, successively joined with other clusters, forming a

dendrogram which can be cut at any height, resulting in different numbers of clusters. (Kaufman & Rousseeuw 2009)

### 4.2.2 Clustering validation

The number of clustering techniques introduced (non-exhaustive) hints at the need to carefully select a method most appropriate to the task at hand. The two main factors to decide on are the clustering technique and the number of clusters. Any number of clusters between 1 and the number of data points is possible, but the question is which number of clusters best explains the inherent structure of the data. To decide which clustering technique to use and to determine the optimal number of clusters, the quality of the result of the clustering has to be measured. Many validation measures exist to meet this requirement and they can be grouped into *internal measures* and *stability measures* but again the question is on which measure one should rely. (Brock et al. 2011)

#### *4.2.2.1 Internal Measures*

Internal measures are only based on intrinsic dataset characteristics to calculate the quality of the cluster. They assess the connectedness (see Connectivity), compactness and separation (combined in the Dunn Index and the Silhouette Width) of the resulting clusters (Brock et al. 2011).

- **Connectivity**

  The Connectivity is calculated using the following equation:

$$Conn(\rho) = \sum_{i=1}^{N} \sum_{j=1}^{L} x_{i,nn_{i(j)}}$$

*Equation 2: Calculation of cluster connectivity (Source: Brock et al. (2011))*

  $nn_{i(j)}$ stands for the $jth$ nearest neighbor of observation $j$, $x_{i,nn_{i(j)}}$ is 0 if $i$ and $j$ are within the same cluster and $1/j$ otherwise. $L$ is a parameter determining how many neighbors to use. The clustering algorithm producing clusters in which most of the data points are grouped together with their nearest neighbors scores a low Connectivity index within the range $[0,\infty]$, which is what should be aimed at.

- **Dunn Index**

  The Dunn index (Dunn, 1974) is calculated using the following equation:

$$D(\rho) = \frac{\min\limits_{C_k, C_l \in \rho, C_k \neq C_l} \left( \min\limits_{i \in C_k, j \in C_l} dist(i,j) \right)}{\max\limits_{C_m \in \rho} diam(C_m)}$$

  *Equation 3: Calculation of the Dunn index of clusters (Source: Brock et al. (2011))*

  $dist(i,j)$ stands for the distance between clusters $i$ and $j$ and $diam(C_m)$ stands for the distance (diameter) within cluster $m$. Both distances can be calculated using a variety of methods, e.g. distance of the most distant elements, distance of centroids or mean distance of all elements. In this thesis, the smallest distance of observations not in the same cluster is compared to the biggest distance between observations within the same cluster. The clustering algorithm producing clusters with a small distance within the lowest density cluster and a big distance between the closest clusters will score a high Dunn index within the range $[0,\infty]$, which is what should be aimed for.

- **Silhouette Width**

  The Silhouette width (Rousseeuw 1987) is calculated using the following equation:

$$S(i) = \frac{b_i - a_i}{\max(b_i, a_i)}$$

  *Equation 4: Calculation of cluster silhouette (Source: Brock et al. (2011))*

  $a_i$ stands for the average dissimilarity of datapoint $i$ with all other datapoints within his cluster and $b_i$ stands for the lowest average dissimilarity of $i$ to any other cluster. The clustering algorithm producing clusters with average $a_i$ significantly smaller than average $b_i$ scores a high Silhouette index within the range [-1,1], which is what should be aimed for.

#### *4.2.2.2 External stability measures*

External stability measures compare the clustering results of clusters obtained from the original datasets to clusters obtained from the original dataset with one measurement removed at a time (Brock et al., 2008). Four measures to assess this stability are proposed:

- **Average Proportion of non-overlap, APN**: The APN measures the average proportion of data points not put in the same cluster in both cases. (Datta & Datta 2003)
- **Average Distance, AD:** The AD measures the average distance between data points put in the same cluster in both cases. (Datta & Datta 2003)

- **Average Distance between Means, ADM:** The ADM measures the average distance between the cluster centroids of data points put in the same cluster in both cases. (Datta & Datta 2003)
- **Figure of Merit, FOM:** The FOM measures the average intra-cluster variance of the deleted data point, considering the clustering result based on the remaining, undeleted observations. (Yeung et al. 2001)

All these measures are averaged over all datasets obtained by removing one observation at a time and they all should be minimized (Brock et al. 2011).

#### *4.2.2.3 Integration of Results*

Table 7 summarizes the presented measures and indicates whether a high or a low value stands for a good clustering result.

| Cluster validation measure | Value indicative of a good clustering result |
|---|---|
| Connectivity | low |
| Dunn-Index | high |
| Silhouette width | high |
| APN | low |
| AD | low |
| ADM | low |
| FOM | low |

*Table 7: Summary of cluster validation measures.*

The different presented internal and stability measures may show different optimal clustering algorithms or numbers of clusters. It is therefore not trivial to decide which clustering algorithm with how many clusters to choose. As a solution to this problem, Pihur & Datta (2007) proposed rank aggregation, which finds, with the Monte Carlo cross entropy algorithm, a *master ranking* that represents the rankings of all the clustering quality measures as good as possible. The winner-algorithm of this master ranking can ultimately be seen as best guess for an optimal clustering algorithm.

#### *4.2.2.4 clValid*

clValid is an R-package supporting the presented cluster validation techniques, addressing the uncertainties concerning selection of clustering technique, determination of optimal number of clusters and selection of validation measures. clValid computes clusters for a given dataset with a defined set of clustering techniques and a defined set of numbers of clusters and produces a defined set of validation measures. The different clustering settings can then be compared and validated on the basis of the different measures, either qualitatively (based on multiple rankings and plots) or quantitatively (based on rank aggregation). (Brock et al. 2011)

## 4.3 CALIBRATION OF THE OPTIMAL MATCHING ALGORITHM AND THE CLUSTERING ALGORITHM

The two main algorithms of this thesis, the sequence alignment optimal matching (OM) algorithm and the clustering algorithm both need to be calibrated before their application. For the OM-algorithm, an

appropriate indel-cost needs to be set and for the clustering algorithm a suitable clustering method and an appropriate number of clusters has to be found. As the evaluation of the results of the OM-algorithm can only be performed based on the resulting clusters of the clustering algorithm, and as the clustering algorithm can only be calibrated with a similarity matrix of the OM-algorithm as input, the calibration of these two analysis steps cannot be performed subsequently. Figure 31 shows the iterative calibration workflow chosen to calibrate the two interdependent algorithms.

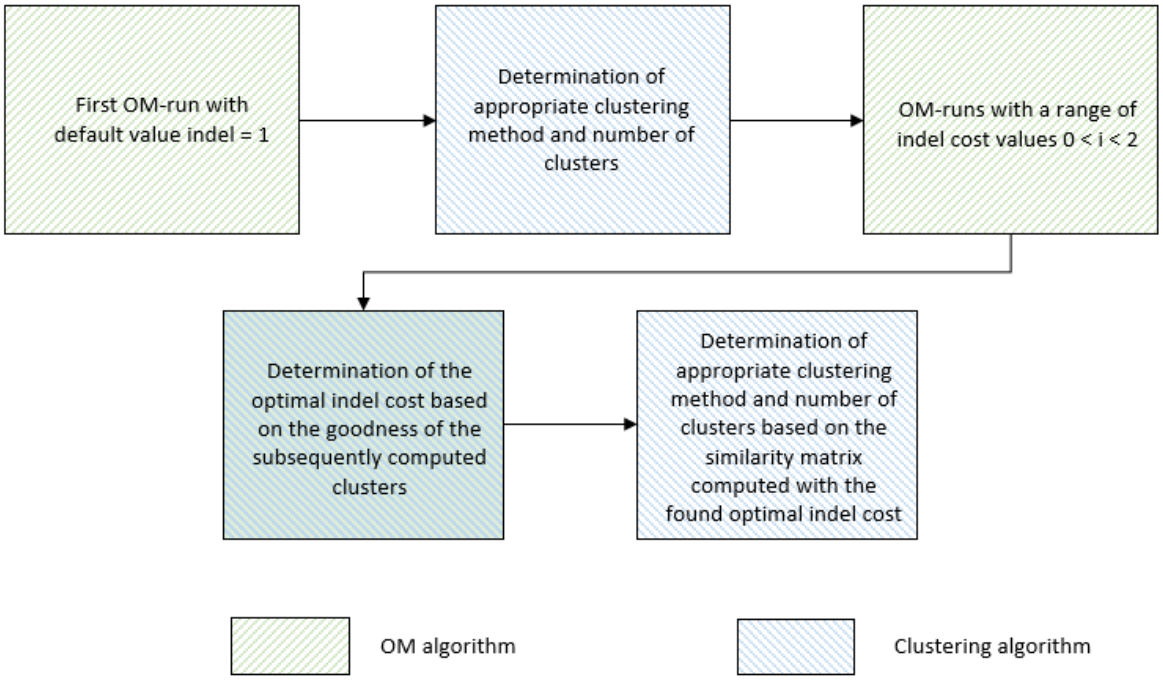
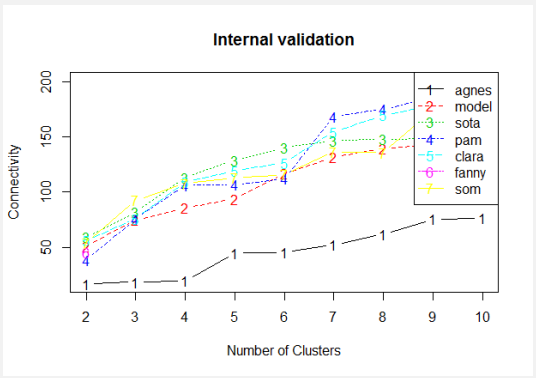


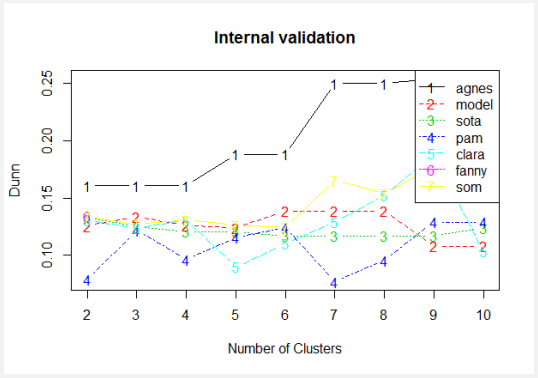

*Figure 31: Calibration workflow.*

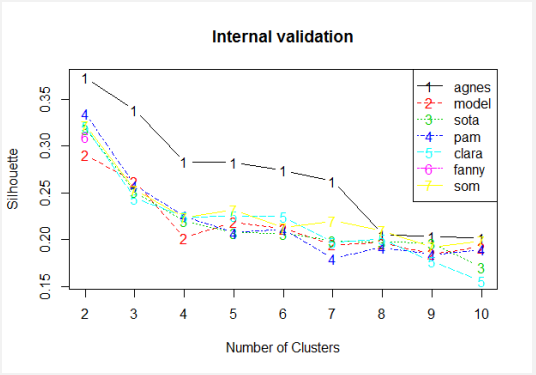
The calibration process was conducted using subset 3 (the smallest subset with 282 trajectories and a maximum gap length of 5 min), as it is the highest-sampled and smallest subset, reducing computational cost. To reduce computation time furhter, the spatial resolution level of floor levels was used. This seems justifiable, as comparisons of the results of the uncalibrated OM-algorithm showed no significant differences for the three scale levels, which was reinforced by the results of the calibrated algorithms (see Section 5.4).

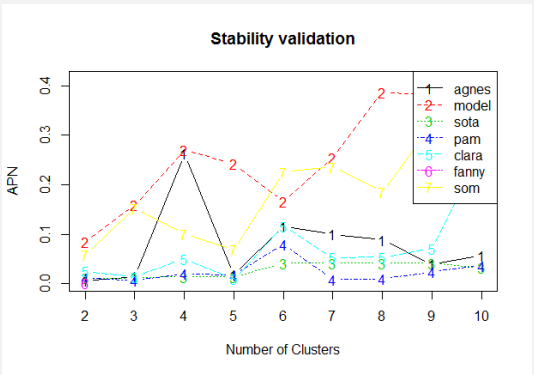
### 4.3.1 Selection of clustering algorithm and number of clusters for indel cost = 1

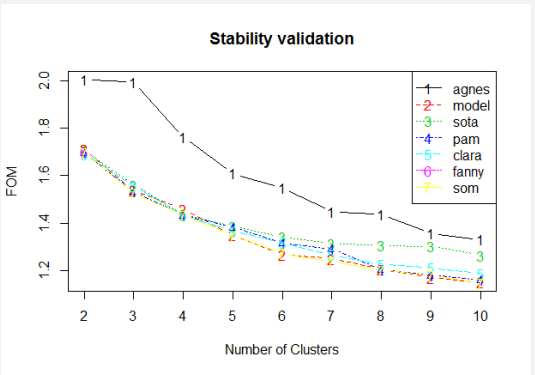
First, a run of the OM-algorithm with an indel cost of 1 was performed, the resulting similarity matrix of which was then used as an input to the clustering algorithm. The clusters of the different clustering methods and with different numbers of clusters were evaluated using internal and stability validation measures (Figure 32).

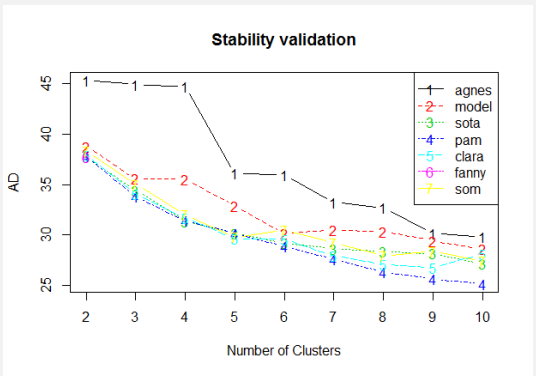


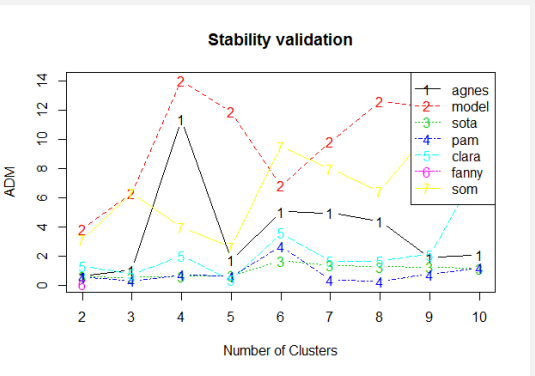




*Figure 32: Internal (Connectivity (low is good), Dunn (high), Silhouette (high)) and stability (APN (low), FOM (low), AD (low), ADM (low)) cluster validation measures for similarity matrix from optimal matching with indel cost = 1.*

The internal validation measures show that AGNES produces the best clusters for basically all numbers of clusters with a clear margin over all the other clustering algorithms. This result matches the proposal of Gabadinho & Ritschard (2009) to use AGNES with TraMineR results. The stability validation measures however show the exact opposite: AGNES performs worst (FOM, AD) or among the worst (APN, ADM). Moreover, the APN and ADM plot reveal a very bad stability for AGNES with 4 clusters.

Using rank aggregation, the values of the different validation measures can be integrated into one ranking. In this case of highly conflicting results, a balance between stability and internal measures has to be found. To this end, following Brock et al. (2011), FOM, which shows similar behavior as AD anyway, was omitted from the ranking. This leaves three internal and three stability measures in the ranking.

Furthermore, unlikely numbers of clusters were omitted, so they would not distort the ranking. If using hierarchical clustering (such as AGNES), a popular way to make a data-informed choice for the number of clusters is to take a look at the dendrogram (Figure 33). The data points (trajectories) at the bottom (height 0) are linked to groups until one big group containing all data points is reached (height 30). The height of all the links is proportional to the dissimilarity of its “daughter-nodes”. The dendrogram can be cut at any height, resulting in different numbers of clusters. Because of the height of all the links being proportional to the dissimilarity of its “daughter-nodes”, the cut is preferably performed at a height were the height difference to the next node is as high as possible. In the presented dendrogram (Figure 33) reasonable choices would be height ≈ {30, 25, 18, 13}, resulting in 2, 3, 5, or 6 clusters. From these clusters the 2-cluster solution was discarded, as a more fine grained segmentation is desired for the later analysis.

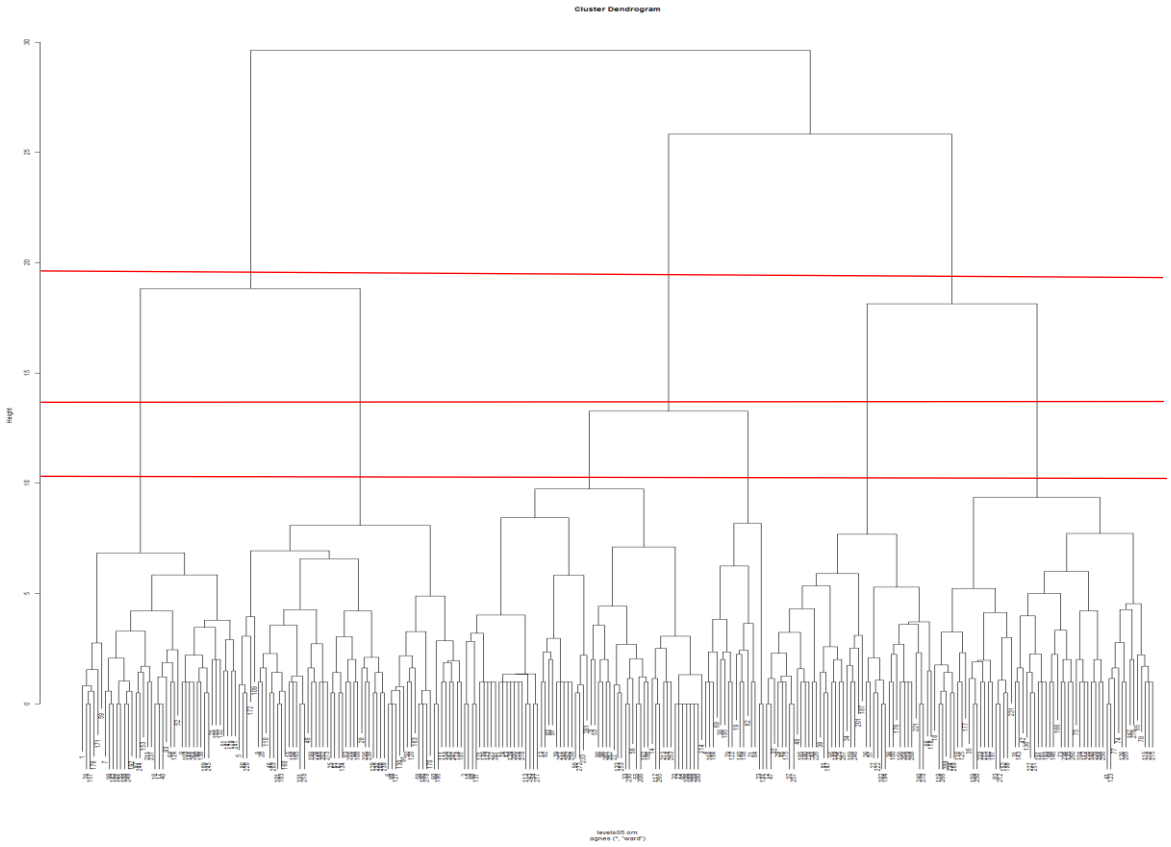



*Figure 33: Dendrogram of the levels distance matrix clustered with AGNES with cut heights for 3,5 and 6 clusters.*

A weighted ranking for six validation measures (Connectivity, Dunn, Silhouette, APN, AD, ADM) of seven clustering algorithms (AGNES, Model, SOTA, PAM, CLARA, FANNY, SOM) and three cluster numbers (3, 5, 6) showed that AGNES with 3 clusters is the most appropriate clustering methodology followed by AGNES with 5 clusters (Figure 34). The grey lines show the ranks of the six validation measures for the different algorithm-number of cluster combinations, the red line shows the best weighted ranking found and can be compared to the black line which shows the mean of the ranks.

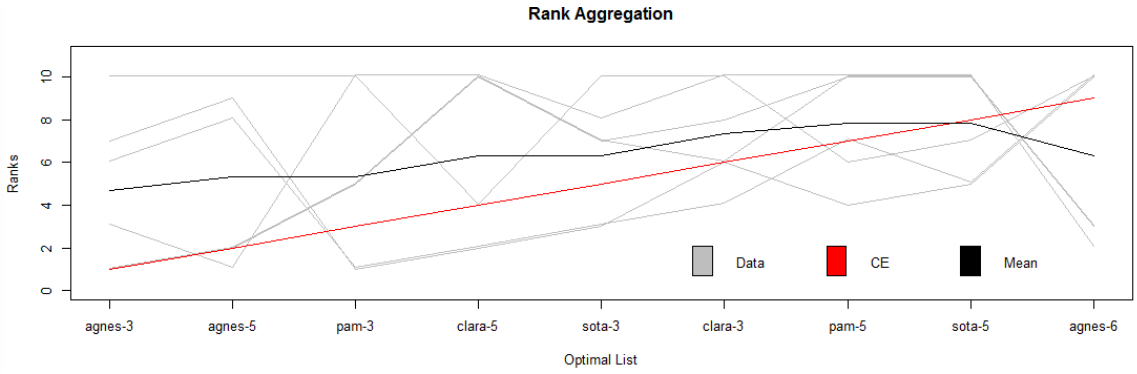



*Figure 34: Ten best-ranked clustering methods with corresponding cluster number for similarity matrix from optimal matching with indel cost = 1.*

### 4.3.2 Selection of indel value for subsequent clustering into 3 clusters using AGNES

To find the indel cost with which the similarity matrix producing is best clusters, the OM-algorithm was run with indel costs = {0.1, 0.5, 1,1.5 ,1.9}. The resulting similarity matrices were clustered using AGNES clustering with 3 clusters. The computed clusters were then validated and the validation results were

integrated using rank aggregation as shown previously. Figure 35 shows that indel cost = 0.5 performed best. It can however be recognized that the difference between the rankings of the different indel values is minimal, as the mean of ranks is close to identical for all costs. In the absence of a better classification methodology, the indel cost 0.5 is chosen for further analysis nevertheless.

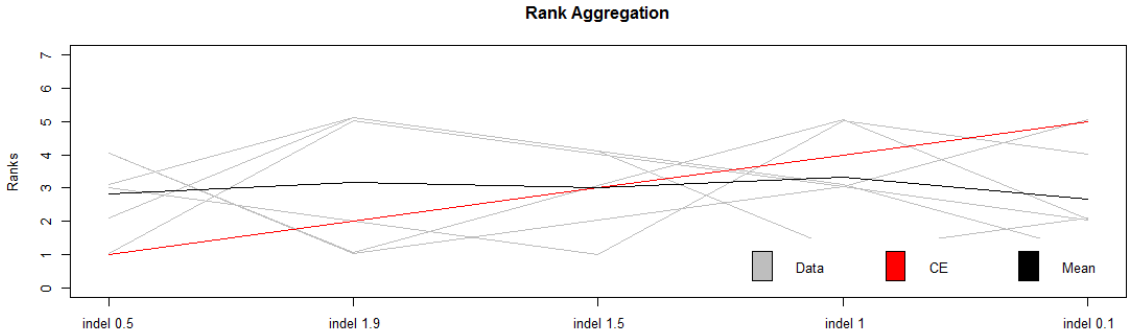



*Figure 35: Ranked indel costs, with which similarity matrices were computed, which were clustered using AGNES and 3 clusters.*

### 4.3.3 Selection of clustering algorithm and number of clusters for indel cost = 0.5

To once again find the most appropriate clustering method, this time with indel cost 0.5 and not just a default value, cluster validation was run for a third time. Figure 36 shows that again AGNES with 3 clusters performed best but this time followed by PAM with 3 clusters and AGNES with 5 clusters in third.

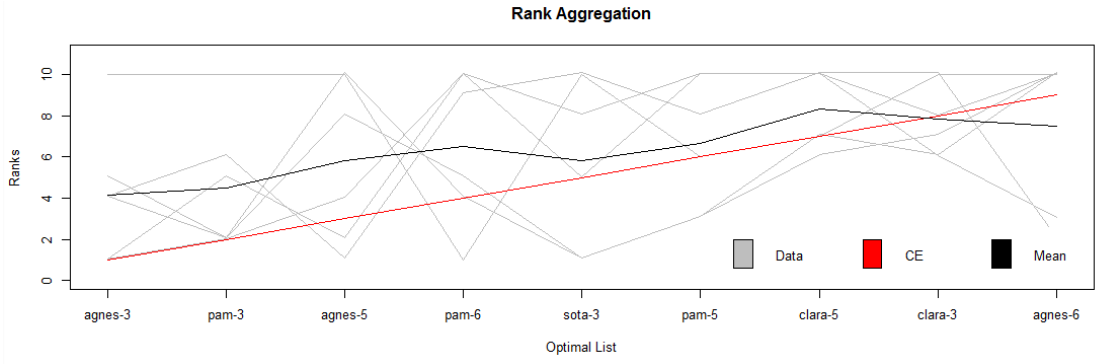



*Figure 36: Ranked clustering methods for similarity matrix from optimal matching with indel cost = 0.5.*

As AGNES with 3 clusters, the final top candidate, coincides with the proposed clustering algorithm by Gabadinho & Ritschard (2009) and as 3 clusters seem like a reasonable choice looking at the dendrogram (Figure 33), the following analysis is performed with this clustering method. As the indel cost of 0.5 produced best clusters, the following analysis is performed with this value.

# 5 RESULTS

In this chapter, the results of the sequence alignment and clustering are presented. Section 5.1 summarizes the sequence characteristics using descriptive plots and statistics, Section 5.2 is about the comparison of intra/inter user sequence similarity, and in Section 5.3 clusters of similar trajectories and their potential relation to the type of shopper are illustrated. All of this is presented for subset 2, the medium dataset, aggregated to building levels. To broaden the scope, Section 5.4 shows the impact of the different spatial and temporal scale levels on the results presented in Sections 5.1-5.3.

## 5.1 SEQUENCE DESCRIPTIVES

Visual inspection of the sequences was used to gain deeper insight into the characteristics of the dataset and to recognize patterns, if possible. The basic characteristics of the sequences are calculated by TraMineR (Gabadinho & Ritschard 2009) and presented here using a sequence plot (all the sequences, for better readability ordered based on sequence similarity, Figure 37), a frequency plot (the 10 most frequent sequences together with their percentage of the whole dataset, Figure 38), a distribution plot (the state distribution at each time step, Figure 39) and a table with the transition rates (state transition distribution, Table 8).

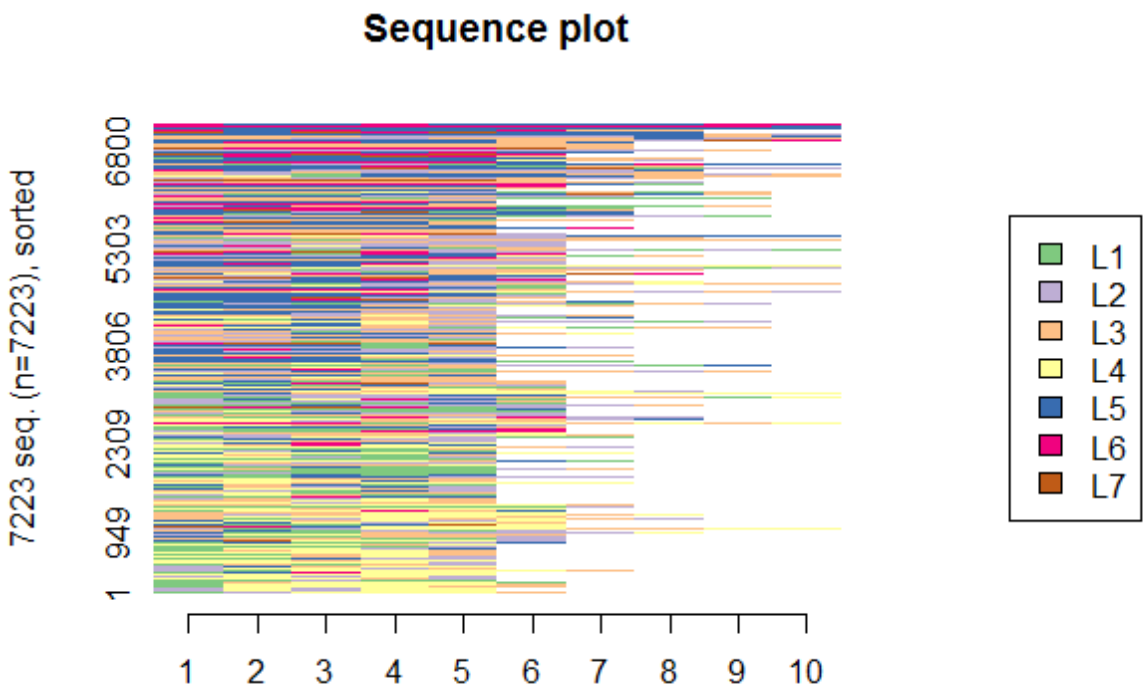



*Figure 37: Sequence plot of subset 2 aggregated to building levels (L1, L2…)showing sequences of fixations (x-axis). Sequences are ordered based on similarity.*

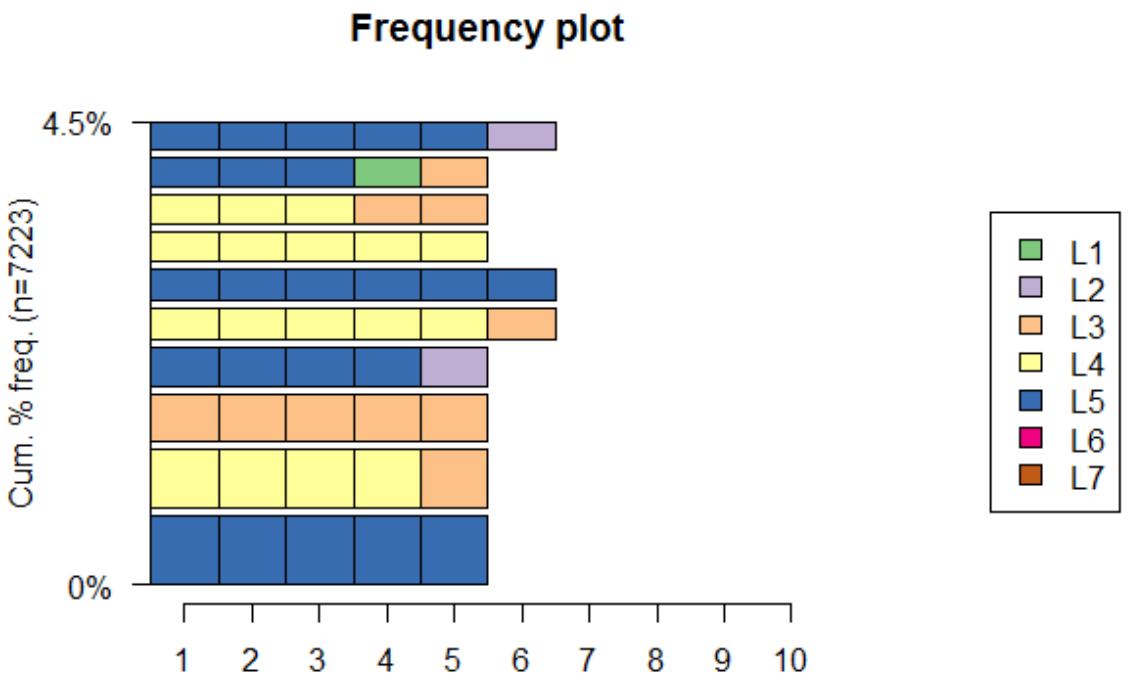



*Figure 38: Frequency plot of the 10 most frequent fixation sequences of subset 2 aggregated to building levels*

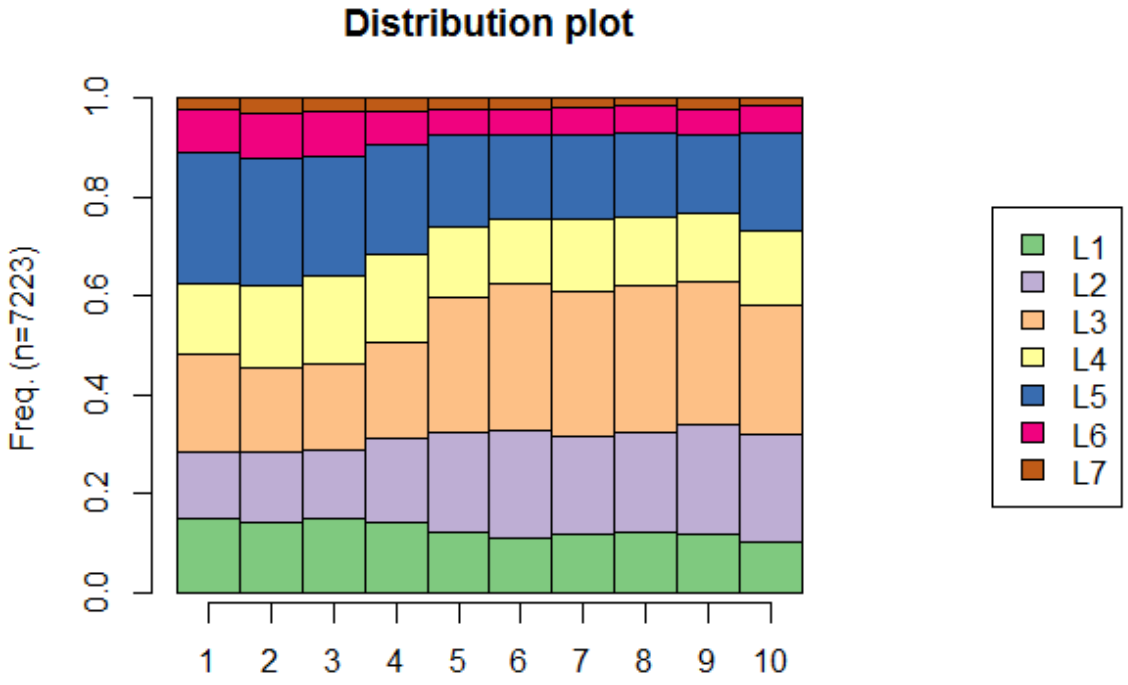



*Figure 39: Distribution plot of subset 2 aggregated to building levels.*

| | [-> 1] | [-> 2] | [-> 3] | [-> 4] | [-> 5] | [-> 6] | [-> 7] |
|---|---|---|---|---|---|---|---|
| **[1 ->]** | 0.34 | 0.24 | 0.16 | 0.06 | 0.17 | 0.02 | 0.01 |
| **[2 ->]** | 0.15 | 0.33 | 0.30 | 0.10 | 0.10 | 0.01 | 0.01 |
| **[3 ->]** | 0.08 | 0.22 | 0.49 | 0.16 | 0.05 | 0.00 | 0.00 |
| **[4 ->]** | 0.04 | 0.10 | 0.27 | 0.54 | 0.04 | 0.00 | 0.00 |
| **[5 ->]** | 0.13 | 0.10 | 0.07 | 0.03 | 0.49 | 0.13 | 0.04 |
| **[6 ->]** | 0.06 | 0.05 | 0.03 | 0.02 | 0.41 | 0.33 | 0.10 |
| **[7 ->]** | 0.09 | 0.11 | 0.05 | 0.02 | 0.28 | 0.26 | 0.19 |

*Table 8: Transition rates of subset 2 (medium dataset) aggregated to building levels as fractions of transitions with each row adding up to 1. Reading example: 24% of the time the state level 1 is followed by the state level 2.*

To better understand the meanings of the different levels of the shopping mall and to be able to identify specific behavior patterns with this knowledge, an overview over the characteristics of the different levels is given. Level 1 is located basically underground and accommodates mainly general grocery and common retail shops as Zara, Adidas etc., as does level 2 which is also the main entrance level. Level 3 and 4 feature more luxurious shops like Armani and Dolce and Gabbana. Level 5 features a big food court, level 6 features a number of more luxurious restaurants and two big Hi-fi shops, while level 7 consists of a viewing platform and a hotel and is only equipped with very few APs.

In the sequence plot (Figure 37) it is hard to detect distinct patterns, as many different sequences exist and there does not seem to be a clearly dominant sequence. Sequences start on all the levels and end on all of them, with very different sequences in between. Nevertheless some patterns are hinted, such as the longer stays on level 5 (food court) that can be detected as a blue bar in the sequence plot or the fact that most of the longer sequences include a stay in the food court.

The frequency plot (Figure 38) supports the presumption that many different movement sequences exist, as the 10 most frequent sequences only account for 4.5% of all the sequences. Most of these most common sequences include a long stay on level 5 and many do not even switch levels. Levels 6 and 7 (luxury restaurant and hotel) are not present in the 10 most common sequences and level 1 only appears once. Level 2 appears twice at the end of the sequence and both times with the pattern L5 -> L2 (food court -> exit level). Another frequent transition is L4 -> L3.

The distribution plot (Figure 39) shows that for time step 1-5 levels 1-4 (different types of shops) all account for 10%-15% of the fixations, while level 5 accounts for approximately 20%, level 6 for 5%-10% and level 7 only for 0%-5%. The distribution for time steps 6-10 is not equally meaningful, as the number of sequences reaching these time steps significantly decreases.

The transition rates (Table 8) show that, with the exception of the rarely visited $6^{th}$ and $7^{th}$ level, most transitions happen within the same level (28%-52%). After these within-level-transitions, transitions to one level above or below are most frequent for levels 1-4 and 6. From level 5 (food court) and level 7 (Hotel) most transitions to another level occur to level 1 and 2 (where most of the entrances are situated).

## 5.2 COMPARISON OF INTRA/INTER-USER SEQUENCES SIMILARITY

The OM-algorithm (used with indel cost 0.5) computes a comprehensive distance matrix with similarity values for each sequence compared to all the others. The mean of all the distances for which both of the sequences originate from the same user (but are not the same sequence) is here called *mean intra-user distance*. Similarly, the mean of all the distances for which the two sequences originate from different users is called *mean inter-user distance*.

| | |
|---|---|
| **Mean intra-user distance** | 3.722 |
| **Mean inter-user distance** | 4.511 |

*Table 9: Mean intra/inter-user distance*

Table 9 shows that the mean intra-user distance is lower than the mean inter-user distance and a student's t-test shows that this difference is significant at the 0.01 level. Thereby the trajectories of a user are more similar to his/her own earlier or later trajectories than to trajectories of other users.

To assess the influence of the indel-cost on this finding, a sensitivity analysis for this parameter with the values {0.1, 0.5, 1, 1.5, 1.9} was conducted. Table 10 compares the intra/inter-user distance from optimal matching distance matrices computed with different indel-costs.

| **Indel-cost** | 0.1 | 0.5 | 1 | 1.5 | 1.9 |
|---|---|---|---|---|---|
| **Intra/inter-user distance** | 0.744/0.902** | 3.722/4.511** | 6.740/8.283** | 8.040/9.654** | 8.868/10.502** |
| **ratio** | 0.825 | 0.825 | 0.813 | 0.832 | 0.844 |

*Table 10: Mean intra/inter-user distance for distance matrices computed with different indel-costs. ** for significant differences (level 0.01).*

All the indel-values lead to significantly different mean intra-user and mean inter-user distances. Also the ratio of the means is stable, with an indel cost of 1 being the optimal cost to detect this difference, but only by a small margin. The fact that the absolute distances for low indel costs are low stems from the many cheap indel operations which are performed, compared to substitutions and more expensive indel operations if indel costs are higher.

## 5.3 CLUSTERS OF SIMILAR SEQUENCES AND THEIR RELATION TO DIFFERENT USER CLASSIFICATIONS

Based on the distance matrix produced by the OM-algorithm, the sequences can be clustered. Figure 40 shows the sequence plot of subset 2 aggregated to building levels, grouped into three clusters computed using AGNES.

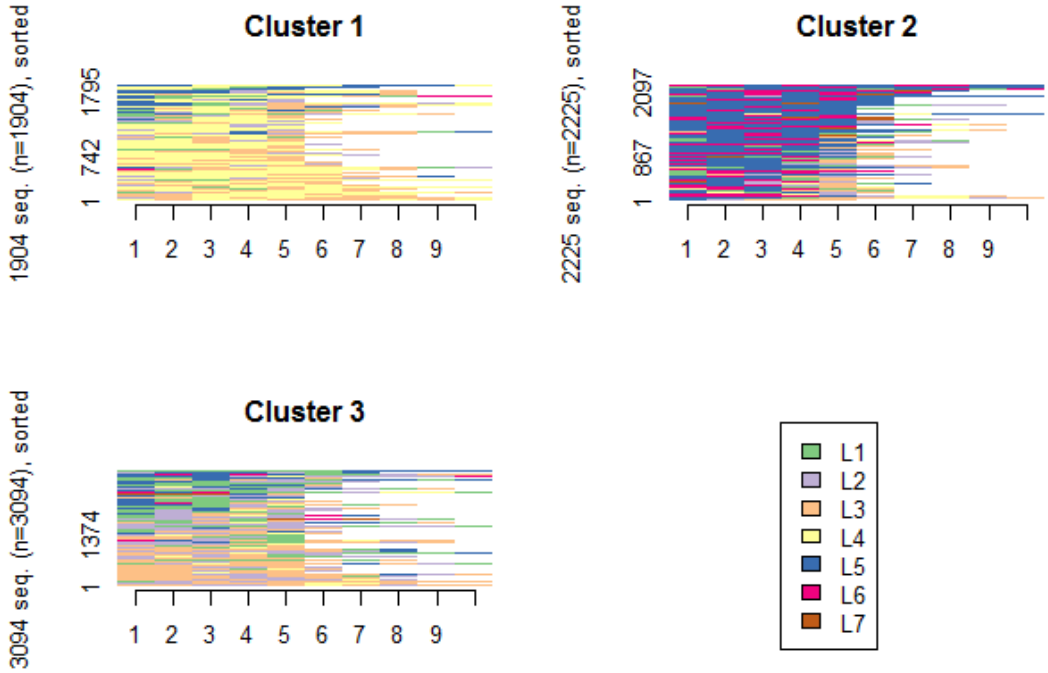



*Figure 40: Sequence plots of the three clusters*

Cluster 1 consists mainly of movement on Level 3 and 4, Cluster 2 of movement on Levels 5-7 and Cluster 3 of movement on Levels 1-3. Most trajectories were assigned to cluster 3 with movement mainly on the entrance and everyday behavior levels, followed by the food court dominated cluster 2 and fewest trajectories were assigned to cluster 1 with movement mainly on the two retail shopping levels. Regarding sequence patterns cluster 1 shows basically all possible patterns of switching between the two levels with a number of sequences starting at the food court level. Cluster 2 exhibits quite stable movements on the two restaurant levels in the beginning of the sequences, with many sequences ending on a lower level. Cluster 3, the biggest cluster, shows the most disturbed sequences, with many switches between the lowest three levels. In Cluster 3, as for Cluster 1, a number of sequences start on the food court level.

Similarly to the intra/inter-user similarity an intra/inter-cluster similarity was computed (Table 11). The intra-cluster similarity has to be lower than the inter-cluster similarity by definition. A student's t-test showed that this difference is significant at the 0.01 level.

| | |
|---|---|
| **Mean intra-cluster distance** | 3.721 |
| **Mean inter-cluster distance** | 4.932 |

*Table 11: Mean intra/inter-cluster distance*

In a subsequent step the relation of the found trajectory-cluster to visitor categories based on temporal characteristics was investigated. Tables 12-15 show the contingency between the three clusters and the four different temporal categorizations presented in Section 3.5.

| | Cluster 1 | Cluster 2 | Cluster 3 | Sum |
|---|---|---|---|---|
| **night (22-6)** | 12 | 27 | 53 | 92 |
| | (13.0%) | (29.3%) | (57.7%) | |
| **morning (6-12)** | 123 | 116 | 210 | 449 |
| | (27.4%) | (25.8%) | (46.8%) | |
| **afternoon (12-18)** | 1505 | 1718 | 2349 | 5572 |
| | (27.0%) | (30.8%) | (42.2%) | |
| **evening (18-22)** | 264 | 364 | 482 | 1110 |
| | (23.8%) | (32.8%) | (43.4%) | |
| **Sum** | 1904 | 2225 | 3094 | 7223 |
| | (26,4%) | (30.8%) | (42.8%) | |

*Table 12: Contingency table for trajectory clusters and user categorization based on time of day with percentage normalized by row.*

| | Cluster 1 | Cluster 2 | Cluster 3 | Sum |
|---|---|---|---|---|
| **weekday (mon-fri)** | 1347 | 1551 | 2186 | 5084 |
| | (26.5%) | (30.5%) | (43.0%) | |
| **weekend (sat-sun)** | 557 | 674 | 908 | 2139 |
| | (26.0%) | (31.5%) | (42.5%) | |
| **Sum** | 1904 | 2225 | 3094 | 7223 |
| | (26,4%) | (30.8%) | (42.8%) | |

*Table 13: Contingency table for trajectory clusters and user categorization based on weekday/weekend with percentage normalized by row.*

| | Cluster 1 | Cluster 2 | Cluster 3 | Sum |
|---|---|---|---|---|
| **Mon** | 250 | 279 | 378 | 907 |
| | (27.6%) | (30.8%) | (41.6%) | |
| **Tue** | 209 | 272 | 355 | 836 |
| | (25.0%) | (32.5%) | (42.5%) | |
| **Wed** | 269 | 292 | 413 | 974 |
| | (27.6%) | (30.0%) | (42.4%) | |
| **Thu** | 364 | 412 | 585 | 1361 |
| | (26.7%) | (30.3%) | (43.0%) | |
| **Fri** | 255 | 296 | 455 | 1006 |
| | (25.3%) | (29.4%) | (45.3%) | |
| **Sat** | 302 | 371 | 507 | 1180 |
| | (25.6%) | (31.4%) | (43.0%) | |
| **Sun** | 255 | 303 | 401 | 959 |
| | (26.6%) | (31.6%) | (41.8%) | |
| **Sum** | 1904 | 2225 | 3094 | 7223 |
| | (26,4%) | (30.8%) | (42.8%) | |

*Table 14: Contingency table for trajectory clusters and user categorization based on weekday with percentage normalized by row.*

| | Cluster 1 | Cluster 2 | Cluster 3 | Sum |
|---|---|---|---|---|
| **< 1 week** | 363 | 602 | 710 | 1675 |
| | (21.7%) | (35.9%) | (42.4%) | |
| **1-2 weeks** | 118 | 131 | 178 | 427 |
| | (27.6%) | (30.7%) | (41.7%) | |
| **2-3 weeks** | 69 | 60 | 56 | 185 |
| | (37.3%) | (32.4%) | (30.3%) | |
| **3-4 weeks** | 24 | 41 | 54 | 119 |
| | (20.2%) | (34.5%) | (45.3%) | |
| **> 4 weeks** | 112 | 123 | 162 | 397 |
| | (28.2%) | (31.0%) | (40.8%) | |
| **Sum** | 686 | 957 | 1160 | 2803 |
| | (24.5%) | (34.1%) | (41.4%) | |

*Table 15: Contingency table for trajectory clusters and user categorization based on return period with percentage normalized by row.*

Whether the distributions shown in the contingency tables showed special patterns and thereby indicated a relation between the categories and the clusters was statistically tested using a Chi-Square test. The test showed a statistically significant (significance level 0.01) relation between the trajectory clusters and the categorization based on time of day as well as based on return period. For time of day it can be observed that when comparing the relative values of cluster membership of the temporal categories to the excepted relative value ( relative value of the sum), disproportionately many visitors at night belong to Cluster 3, disproportionately few in the morning belong to Cluster 2 and disproportionately few in the evening to Cluster 1 (Table 16). An example for return period would be that disproportionately few visitors with a return period <1 week belong to cluster 1 but that disproportionately many with a return period of 2-3 weeks do (Table 17).

| | Cluster 1 | Cluster 2 | Cluster 3 |
|---|---|---|---|
| **night (22-6)** | 0.49 | 0.95 | 1.32 |
| **morning (6-12)** | 1.03 | 0.83 | 1.08 |
| **afternoon (12-18)** | 1.02 | 1.00 | 0.97 |
| **evening (18-22)** | 0.90 | 1.06 | 1.00 |

*Table 16: Quotient of relative cluster membership and expected relative cluster membership (sum) for user categorization based on time of day.*

| | Cluster 1 | Cluster 2 | Cluster 3 |
|---|---|---|---|
| **< 1 week** | 0.88 | 1.05 | 1.02 |
| **1-2 weeks** | 1.12 | 0.90 | 1.00 |
| **2-3 weeks** | 1.52 | 0.95 | 0.73 |
| **3-4 weeks** | 0.82 | 1.01 | 1.09 |
| **> 4 weeks** | 1.15 | 0.90 | 0.98 |

*Table 17: Quotient of relative cluster membership and expected relative cluster membership (sum) for user categorization based on return period.*

No significant relationship could be detected between the trajectory clusters and the categorization based on weekday/weekend or based on weekday.

## 5.4 CROSS-SCALE-ANALYSIS FOR TIME AND SPACE

The results presented in the previous Section were produced only in respect to one temporal scale level (subset 2, gap length/ temporal resolution <=7.5 min) and one spatial scale (spatial aggregation to building levels). These scale levels were deliberately chosen as the data were selected based on temporal resolution and were then aggregated to the desired spatial resolution. The goal of the cross-scale analysis presented in this section is to assess the effect of changing spatial and temporal resolution on the presented results.

### 5.4.1 Comparison of intra/inter-user sequences similarity

The mean intra-user distance was found to be significantly lower than the mean inter-user distance for subset 2 aggregated to building levels. Table 18 shows these measures for all scale combinations and indicates with stars which of these means are significantly different (significance level 0.01).

| | hotspots | areas | levels |
|---|---|---|---|
| **Subset 1** | 5.758/6.243** | 5.095/5.782** | 3.721/4.709** |
| **Subset 2** | 5.544/6.001** | 4.923/5.552** | 3.722/4.511** |
| **Subset 3** | 5.559/4.944 | 5.147/4.569 | 4.029/3.705 |

*Table 18: Mean intra/inter-user distance dependent on subset size (temporal resolution) and aggregation level (spatial resolution). ** for significant differences (level 0.01).*

Subset 3, the smallest subset, shows higher intra-user than inter-user distance, however the differences are not significant. Subset 1 and 2 show lower intra-user than inter-user distance and these differences are significant. Furthermore, the differences in intra- and inter-user distance mean between subset 1 and 2 for all the spatial scale levels are very small.

### 5.4.2 Comparison of clusters of similar sequences and their relation to different user classifications

Three clusters of similar trajectories were found in Section 5.3. Table 19 compares clusters, obtained by clustering distance matrices from different spatial resolutions, to the original clustering result. The measures given include the adjusted Rand index (a measure comparing the number of agreements to the number of disagreements of the two clustering results; Rand 1971), average within/between cluster distance and the Dunn index.

| | Subset 2 |
|---|---|
| **hotspots** | **hotspots** |
| | **levels** / **1** / **2** / **3**<br>**1**: 969 / 781 / 154<br>**2**: 10 / 253 / 1962<br>**3**: 73 / 2201 / 820<br><br>Adjusted Rand index = 0.327<br>Avg_within_cluster distance = 5.787<br>Avg_between_clusters distance = 6.135<br>Dunn index = 0.1 |
| **areas** | **areas** |
| | **levels** / **1** / **2** / **3**<br>**1**: 1249 / 528 / 127<br>**2**: 14 / 255 / 1956<br>**3**: 105 / 2579 / 410<br><br>Adjusted Rand index = 0.484<br>Avg_within_cluster distance= 5.155<br>Avg_between_clusters distance= 5.787<br>Dunn index = 0.1 |
| **levels** | Avg_within_cluster distance= 3.721<br>Avg_between_clusters distance= 4.932<br>Dunn index = 0.05 |

*Table 19: Contingency table and adjusted Rand index for hotspots and areas compared to levels and Dunn Index for all the scale levels*

Both clustering results, based on hotspots and areas, show similarities to the levels clustering results (clusters 2 and 3 simply swap places), however the relation of the areas-clustering to the levels-clustering is higher than the one of the hotspot-clustering (higher adjusted Rand index).

The new clustering results based on areas and hotspots were also tested for a relation to the temporal visitor categories. Whether the different clustering results significantly relate to the temporal user categories is summarized in Table 20.

| | Time of day | Weekday/weekend | Weekday | Return period |
|---|---|---|---|---|
| **Hotspots** | 0.0005** | 0.6176 | 0.1671 | 0.0003** |
| **Areas** | 0.0040** | 0.8574 | 0.8627 | 0.0017** |
| **Levels** | 0.0017** | 0.6995 | 0.8893 | 0.0001** |

*Table 20: Relationship between clustering results based on different scale levels and different temporal user categorizations using p-value. **for significant values (level 0.01).*

The significant relations found between clusters and user categorization are the same for all the spatial scale levels. However, clusters from clustering based on hotspots show the strongest relation to the temporal user categorization with the exception of return period.

# 6 DISCUSSION

This chapter addresses the research questions stated in the beginning of this thesis and discusses to what extent the chosen methodology was helpful in answering them. Section 6.1 discusses whether recurrent patterns of behavior were found and how this can be interpreted (RQ1). In Section 6.2, the relation between the clusters of similar trajectories and the different temporal user-categories is reviewed (RQ2). Section 6.3 discusses the executed multi-scale analysis in respect to the effect of spatial and temporal resolution on the before discussed results (RQ3), and lastly Section 6.4 evaluates the chosen sequence alignment method based on its validity, credibility and efficiency.

## 6.1 RECURRENT PATTERNS OF BEHAVIOR

The optimal matching distance between trajectories of the same user was found to be significantly smaller than the one between trajectories of different users. As this result showed to be stable independent of the indel cost, hypothesis 1, that *symbolic representation of the trajectories can be applied in combination with Sequence alignment methods (SAM) to identify recurrent movement patterns in Eulerian datasets of indoor movement*, can be accepted.

The fact that recurrent shopping behavior exists is broadly accepted in the research community (Sheth & Raju 1974; Laaksonen 1993; Mulligan 1987). Sheth & Raju (1974) identify four basic purchase behaviors: Habitual, exploratory, impulsive and belief based behavior. Both habitual behavior, going to the shop one has always gone to, and belief based behavior, going to the shop that, according to one's beliefs, sells the best products, explain the repetitive character of shopping behavior. Laaksonen (1993) sees the main reason for recurrent shopping behavior in customer loyalty. According to Sheth & Park (1974) this loyalty can grow with or without considering evaluative or emotional aspects, translating nicely to the concepts of habitual and belief base behavior presented by Sheth & Raju (1974). To differentiate between these two causes for recurrent behavior, Laaksonen (1993) proposes to test whether a customer shows resistance to persuasion to switch. If for example a customer's favourite shop closes and the customer then drives to the same shop in the next city instead of just going to another shop in his hometown, belief based behavior can be observed rather than habitual behavior.

The confidence in the repetitiveness of shopping behavior is big enough to build entire behavior and decision models on it. Mulligan (1987) for example designed a multipurpose shopping model with the goal of predicting whole shopping tours, depending on what items are needed.

In the light of this research on recurrent shopping behavior, the found repetitive movements in the shopping center are sufficiently explained and could be expected, as a customer is likely to move in a recurrent manner through the mall if he visits nearly the same shops during all his visits. However, to my knowledge, this hypothesis has up to the analysis presented in this thesis not been confirmed by pure trajectory analysis on such a large dataset.

## 6.2 RELATION BETWEEN CLUSTERS OF SIMILAR TRAJECTORIES AND TYPES OF USERS

Similar groups of people showing similar shopping behavior is a broad consensus in the research field of marketing (e.g. Zeithaml 1985; Wesley et al. 2006). Zeithaml (1985) identified the five demographic variables sex, female working status, age, income and marital status to significantly influence shopping behavior. Wesley et al. (2006) confirmed that also twenty years later, gender still is the most important

descriptive of shopping behavior. Lee Taylor & Cosenza (2002) showed that the user groups do not need to be that broad to show distinct behavior, when they found and investigated characteristic shopping behavior of the *later aged female teens*. That not only gender and status related descriptives have an influence on the shopper typification was shown for example by Stafford et al. (2014) when they found distinct online shopping behavior for different cultural groups, specifically shoppers from the United States, Finland and Turkey. To bring the scope back to physical shopping in a shopping mall, for which also movement can be analyzed, Jarboe & McDaniel (1987) linked demographic variables also to browsing behavior, a specific shopping behavior including a lot of movement by definition.

In this thesis, groups of similar behavior have been detected in respect to the similarity of their movements. Whether these groups can be explained with demographic variables as presented in the previous paragraph could, as explained before, not be established due to lack of demographic data. However, temporal characteristics of a user's trajectories could be used to derive a number of user categorizations, two of them, time of day and return period, with explaining power regarding groups with similar behavior.

Users most often present in the shopping center at night, and thereby outside the mall's opening hours, can be expected to be mall employees. Most of the trajectories of these users were assigned to cluster 3, the cluster with most turbulent and mixed sequences, which indicates, that the employees (e.g. cleaning staff, security) moved through the whole mall. In the morning, normal shopping behavior on the everyday needs levels dominates while in the evening the food court and the restaurants are well frequented.

Regarding the return period the most striking finding is that trajectories of visitors who visit the shopping center more than once a week in average, are not likely to be assigned to the cluster mainly containing movement on the retail levels. This can be explained by the fact that new clothes or sport equipment do not have to be bought every week. Visitors returning more than once a week therefore mainly move through the levels with everyday products (an example could be food shopping) or they go to the food court/restaurants (an example could be nearby office employees eating lunch in the mall).

The day of week users usually go shopping did not show to have an influence on the users' indoor movement behavior. As far as one can tell from analyzing the trajectories, visitors coming to the mall on weekends seem to have similar needs as visitors coming during the week and the same counts for visitors coming on the Australian shopping day, Thursday, compared to visitors coming on the calmest day, Tuesday.

To summarize, hypothesis 2, that *the type of shoppers has a significant relationship with the trajectory patterns and shoppers with similar characteristics are expected to have similar trajectories*, can be accepted for user categorization based on time of day and return period, but can neither be accepted for user categorization based on day of week nor based on weekday/weekend.

## 6.3 EFFECT OF SPATIAL AND TEMPORAL RESOLUTION ON THE RESULTS

The influence of spatial and temporal granularity on the results of sequence alignment methods is one of the big uncertainties of the method (Shoval et al. 2015). In the light of Levin (1992) stating that what patterns researchers find is dependent on the scale they look at a phenomenon, it could be expected

that looking at the movement traces at different spatial and temporal granularities allows identifying different patterns. The cross-scale analysis performed in this thesis examines potential effects of changing granularities on the identification of recurrent movement patterns and on the formation of groups showing similar movements.

In respect to recurrent patterns of behavior detectable, temporal scale at first seems to have an influence on the results as intra-user distance showed to be bigger than inter-user distance for subset 1. However, subset 1, only featuring 6 users coming to the mall more than once can be seen as not admissible for this kind of analysis as the outcome is likely pure chance. As similarly strong recurrent patterns could be detected for subset 2 and subset 3, the best guess is that rather a critical subset size with enough returning users has to be reached than an influence of temporal granularity of the trajectories can be argued. Regarding spatial granularity, recurrent patterns could be detected totally independently of the spatial scale. This means that characteristic routes of shoppers returning to the mall several times a year do not only manifest themselves through micro route choices within a level but also through the sequences of visited levels.

Groups of similar trajectories computed at the spatial granularity of hotspots and areas were similar to the groups computed at the granularity of levels and clusters were of similar quality regarding the Dunn-Index, with level-granularity scoring slightly worse than the other two. The relation to the types of users again seemed to be close to identical for all scale levels, which leads to the conclusion that finer spatial granularity leads to slightly better clusters, but the different cluster do not differ too much from each other and the relation to user categories does not differ at all.

A scale or dimension not investigated in the scope of this thesis is the functional dimension. Space could be categorized into functional parts, which would allow to test specific hypotheses regarding the behavior of the visitors. The TRIIBE team for example used categorization into navigational, retail and food court context to investigate the influence of space on visitors' information behavior, but also finer categorizations, for example based on type of shop are imaginable.

The argumentation presented in this section leads to the conclusion that hypothesis 3, *trajectories expressed at coarser spatial and temporal granularities are less specific and distinguishing than if analyzed at finer granularities and that a granularity as fine as possible produces best results for RQ1 and RQ 2*, cannot be accepted. This by no means proves Levin (1992) wrong, it just shows that the two investigated patterns, if using SAM and clustering methods, can be found basically independent of spatial and temporal granularity. This again supports the choice of SAM as an appropriate movement mining tool for such kind of data in the first place. Finer spatial and temporal resolution combined with a more *geometry-based* movement mining tool may give insight into different behavioral patterns as for example micro route choices, but for analyzing the given sparsely-sampled dataset SAM proved to be a sensible choice, as even coarser granularities only slightly affect the findings. It can therefore be argued that this cross-scale analysis only compared *coarse* and *even coarser* granularities and that finer scale, temporal as well as spatial, may have indeed an influence on the SAM results.

## 6.4 EVALUATION OF SEQUENCE ALIGNMENT AS A TECHNIQUE FOR MOVEMENT MINING IN A SPARSELY SAMPLED EULERIAN MOVEMENT DATASET

The application of SAM with sparsely sampled Eulerian trajectories requires accurate pre-processing (Section 3.4), careful calibration of the optimal matching algorithm and the subsequently used clustering algorithm (Section 4.3) and thorough evaluation of the results concerning validity (Section 6.4.1), credibility and interestingness (Section 6.4.2) and efficiency (Section 6.4.3) (evaluation following Laube (2014)).

### 6.4.1 Validation

A valid method shows an acceptable range of accuracy compared to the intended goal of the method (Laube 2014). One way to test the validity of a method is a sensitivity analysis for the parameter values. The sensitivity of the results to indel cost, clustering algorithm and number of clusters was not in every case tested explicitly. Nevertheless the parameters were carefully calibrated and the comparison of effects of different indel costs on cluster quality for example shows, that different values perform well regarding different evaluation measures and no clear best value can be found. This leads to the expectation that the sensitivity of this specific parameter is not critical.

Another possible validation strategy is to visualize the results and compare the graphs to expected results. As trajectories that could visually be recognized as very similar were grouped in the same cluster (Figure 40) the results can be qualified as valid from this point of view.

Other ways of testing the validity such as expert interviews or comparisons to established methods were not included in the scope of this thesis, the latter also partly due to the fact that no broadly established movement mining method for this kind of data exists. However, outcomes of further validation in the mentioned ways are not expected to be negative.

### 6.4.2 Credibility and Interestingness

Credibility with the notion of interestingness as Laube (2014) uses it, evaluates the backup of domain experts of a method, and thereby also assesses the usefulness and applicability of the results in its real life context. This thesis, as a part of the TRIIBE project, works closely together with the data providers (the shopping mall owners) so the results are meant to be interesting to them. Following Silberschatz & Tuzhilin (1996) interestingness is composed of unexpectedness and actionability of the results. While the acceptance of Hypotheses 1 may not be unexpected, it supports actions and decisions of shop and mall owners, for example in the form of indoor recommender systems. Recommender systems collect data from a variety of sources to make as well-informed and personalized recommendations as possible, the recommendations reaching from where to go next, to what to do or buy or who to interact with next (Adomavicius & Tuzhilin 2005). Recommender systems acquire their information from every possible aspect of a user's online interaction (Ricci et al. 2011) and if position and movement of the user are part of this interaction they can be used for more accurate recommendations. If a user's last stop of a shopping trip often is a beer at the same bar, an intelligent recommender system could make such recommendations when it realizes that the shopping trip is about to come to an end.

The findings that some user groups (in this thesis temporal user groups) have a relation to the way they move through the shopping center (Hypothesis 2) can have some level of unexpectedness and has actionability in the domain of recommender systems too. Recommender systems not only rely on

personal historic information but are also knowledge-based (most people like coffee in the morning), community-based (my friend likes action movies so I probably like them too) and demography-based (my cohort of wealthy, suburban, middle aged women likes the luxury stores on the sixth level, so I do probably too) (Ricci et al. 2011). Although the user groups in this thesis were not demographic, they were planned to be demographic in the beginning and could be in a future study and in such a case could support the demography-based part of recommender systems. The thesis as is can at least consolidate existing knowledge such as the fact that shopping mall restaurants are frequented most often around lunch- and dinnertime.

The results concerning the influence of spatial and temporal scale are of interest to the shopping mall owners as it relates to the cost of the infrastructure (tracking method, data storage) and as location tracking, as discussed in section 2.1.3, comes with privacy issues, so the lowest resolution that still holds interesting information is to be preferred. Moreover, these results are of interest to the academic community (Shoval et al. 2015) and can strengthen the credibility of the other results by eliminating unvertainties about the method.

### 6.4.3 Efficiency

The performance of the algorithms of a method define its efficiency and thereby also its scalability (Laube 2014).

The complexity of sequence alignments grows with the square of the number of sequences ($O(n^2)$) and is furthermore dependent on the length of the sequences (Dodge et al. 2012). However, this does not seem to be too much of a problem for most researchers of which studies were discussed in this paper, as neither the designer of the base software (Thompson et al. 1994), nor the designers of the software applied in the spatial domain (Clustal G/TraMineR) (Wilson et al. 1999; Wilson 2001; Gabadinho & Ritschard 2009) nor the users of the software (e.g. Shoval & Isaacson 2007; De Groeve et al. 2015) discuss it. This leads to the conclusion, that these researchers either used powerful computers (compared to the off-the-shelve computer used for this thesis (Asus UX305F, Intel Core M-5Y10, 8GB RAM), or analyzed only small datasets. In any case, the sequence alignment of all 260'296 sequences of the original dataset using TraMineR would overcharge the computer used to compute the analysis presented in this thesis heavily, as expected computation time would be 1.5 hours (still manageable) and the needed memory space, as R stores objects in memory (Gabadinho & Ritschard 2009), would amount to 510 Gb (both values extrapolated).

The chosen clustering algorithm, AGNES, normally has a complexity of ($O(n^3)$) which makes it hard to use with very large datasets (Kaufman & Rousseeuw 2009). The calibration workflow, presented in Section 4.3, included cluster computation after removal of one observation at a time, resulting in a complexity of ($O(n^4)$). The subset for the calibration therefore has to be reasonably small (subset 3) and also one single cluster computation becomes very expensive with growing subset size. Clustering subset 1 took approx. 4.5 hours, extrapolating this on the whole dataset would result in 223 hours.

Summarizing this section, it has to be admitted that neither the sequence alignment nor the clustering algorithm show sufficiently efficient behavior to claim that scalability is ensured. This elucidates the importance of finding meaningful subsets that allow to translate the findings to the bigger scale, which cannot be analyzed directly.

# 7 CONCLUSION

A summary of the analysis and the findings (Section 7.1), an overview of the contributions of this thesis (Section 7.2) and an outlook for future work (Section 7.3) is given in respect to the research field of sequence alignment in movement analysis as well as in respect to this thesis being a part of the TRIIBE project.

## 7.1 SUMMARY

The analysis of people's movements, being the analysis of a fairly direct trace of human behavior, is not only an inherent part of the research field of GIS, but also enjoys high acceptance and frequent use in the social sciences as for example marketing or behavioral studies. Movement data from various data sources may be mined for distinct patterns, this thesis having analyzed an extensive Wi-Fi tracking dataset from a shopping mall. The dataset, coming with important advantages as enabling movement tracking in indoor environment or the large population recorded, imposed challenges regarding the choice of appropriate movement mining tools due to its low spatial and temporal resolution. SAM proved to be a valid choice for analyzing data like this, as it enabled the researchers to elucidate expected patterns, as individual recurrent behavior, and to find clusters of similar movements through the shopping mall and assign these groups to different types of users, all of this solely based on trajectory analysis. Furthermore, a cross-scale analysis investigating the influence of spatial and temporal granularity on the identified patterns and groups showed that, if SAM on the Wi-Fi tracking dataset is used, different spatial and temporal resolutions only have a small effect on the results. Finally, the applied methods were evaluated as valid as well as credible and interesting, with some persisting challenges regarding their efficiency and scalability.

## 7.2 CONTRIBUTIONS

This is the first attempt of applying SAM to a Wi-Fi tracking dataset to mine distinct movement patterns and it is the first implementation of SAM in a shopping mall environment. Furthermore, the low instrumentalisation of the study setup as well as the large size of the dataset distinguish the analysis performed from most of the previous studies and come with opportunities as well as challenges. This thesis demonstrates how carefully calibrated sequence alignment and clustering algorithms applied to a Wi-Fi tracking dataset of a big shopping mall can be used to investigate and prove behavioral patterns (repetitive movements, similar movements of similar types of shoppers) suggested by behavioral and marketing studies.

A second contribution of this thesis is the analysis of the effects of changing spatial and temporal resolution on the sequence alignment results, a research gap pointed out by Shoval et al. (2015). The found small influence of spatial and temporal resolution on the patterns and groups identified in this thesis is not directly transferable to other datasets and/or research questions, nevertheless this thesis presents a first analysis of this important aspect of SAM.

## 7.3 OUTLOOK

SAM is a well established method in tourism research (Bargeman et al. 2002; Lee & Joh 2010; Shoval et al. 2015) and there is no obvious reason for it not to be the same in a shopping environment. Future Wi-Fi tracking datasets are expected to feature more than one AP at a time, allowing for a significantly finer spatial tracking resolution that in turn allows detection of different patterns. Movement patterns such as route choice or a shopper typology similar to the one of Lykourentzou et al. (2013) or Kuflik &

Dim (2013) comparing individual (Lykourentzou et al. 2013) and social (Kuflik & Dim 2013) movement patterns of museum visitors to typical animal behavior (Figure 41) potentially could be detected with such data.

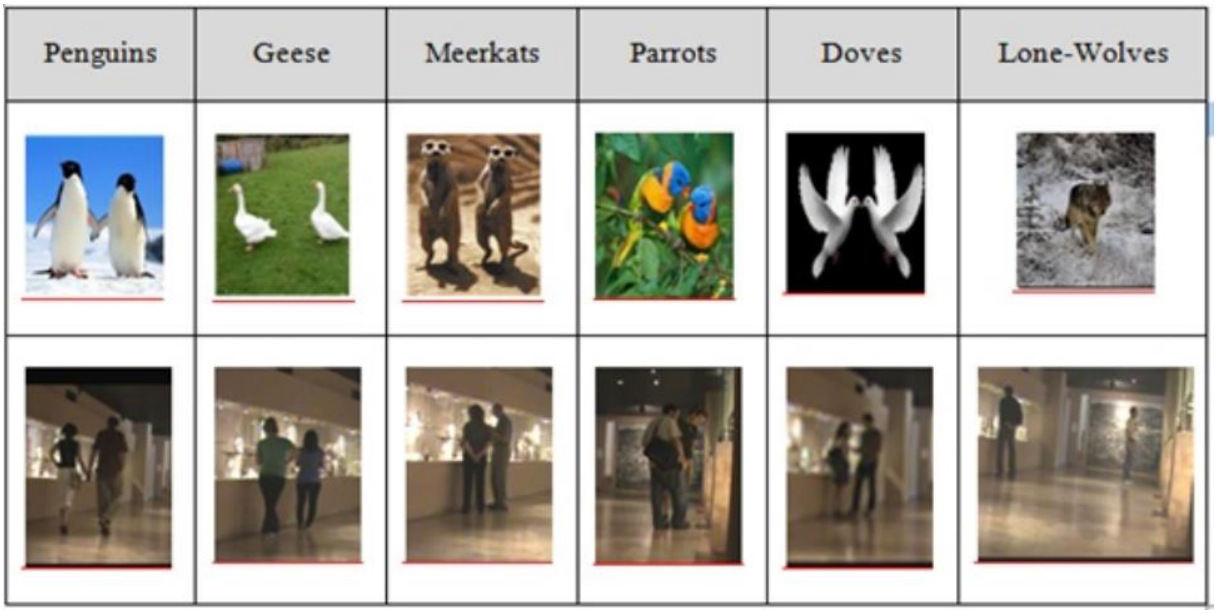



*Figure 41: Visitors pair-behavior types compared to typical animal behavior. Source:* Kuflik & Dim (2013)

Regarding the dataset as it is now, future work could involve the implementation of a prediction model as presented by Hawelka et al. (2015) to predict the future movements of a visitor based on his/her previous movements or to fill gaps in trajectories based on his/her previous movements or movements of other visitors. Another potentially interesting future research step of the TRIIBE project could be to link the found movement patterns and groups to the web behavior of the visitors, to investigate the relationship of indoor movement and information behavior.

# REFERENCES


Abbott, A. & Forrest, J., 1986. Optimal matching methods for historical sequences. *The Journal of Interdisciplinary History*, 16(3), pp.471–494.

Adomavicius, G. & Tuzhilin, A., 2005. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6), pp.734–749.

Andrienko, N. et al., 2008. Basic concepts of movement data. In *Mobility, data mining and privacy*. pp. 15–38.

Andrienko, N. & Andrienko, G., 2007. Designing Visual Analytics Methods for Massive Collections of Movement Data. *Cartographica: The International Journal for Geographic Information and Geovisualization*.

Bai, Y.B. et al., 2014. A New Approach for Indoor Customer Tracking Based on a Single Wi-Fi Connection. In *2014 International Conference on Indoor Positioning and Indoor Navigation*.

Bargeman, B., Joh, C.-H. & Timmermans, H., 2002. Vacation behavior using a sequence alignment method. *Annals of Tourism Research*, 29(2), pp.320–337.

Bell, S., Jung, W. & Krishnakumar, V., 2010. WiFi-based enhanced positioning systems: accuracy through mapping, calibration, and classification. *Proceedings of the 2nd ACM*, pp.3–9.

Bleisch, S. et al., 2013. Mining candidate causal relationships in movement patterns. *International Journal of Geographical Information Science*.

Both, A. et al., 2012. Decentralized Monitoring of Moving Objects in a Transportation Network Augmented with Checkpoints. *The Computer Journal*, 56(12), pp.1432–1449.

Brock, G., Pihur, V. & Datta, S., 2011. clValid, an R package for cluster validation. *Journal of Statistical Software*, 25, pp.1–22.

Buchin, K. et al., 2011. Finding long and similar parts of trajectories. *Computational Geometry*, 44(9), pp.465–476.

Chan, T., 1995. Optimal matching analysis: a methodological note on studying career mobility. *Work and occupations*, 22(4), pp.467–490.

Cheng, R. et al., 2006. Preserving user location privacy in mobile data management infrastructures. *Int. Workshop on Privacy Enhancing Technologies*, pp.393–412.

Çöltekin, a., Fabrikant, S.I. & Lacayo, M., 2010. Exploring the efficiency of users' visual analytics strategies based on sequence analysis of eye movement recordings. *International Journal of Geographical Information Science*, 24(10), pp.1559–1575.

Datta, S. & Datta, S., 2003. Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics*, 19(4), pp.459–466.

Delafontaine, M. et al., 2012. Analysing spatiotemporal sequences in Bluetooth tracking data. *Applied Geography*, 34, pp.659–668.

Dodge, S., Laube, P. & Weibel, R., 2012. Movement similarity assessment using symbolic representation of trajectories. *International Journal of Geographical Information Science*, 26(9), pp.1563–1588.

Dodge, S., Weibel, R. & Lautenschütz, A.-K., 2008. Towards a taxonomy of movement patterns. *Information Visualization*, 7(3-4), pp.240–252.

References


Dopazo, J. & Carazo, J.M., 1997. Phylogenetic Reconstruction Using an Unsupervised Growing Neural Network That Adopts the Topology of a Phylogenetic Tree. *Journal of Molecular Evolution*, 44(2), pp.226–233.

Duckham, M. & Kulik, L., 2005. *Pervasive Computing* H.-W. Gellersen, R. Want, & A. Schmidt, eds., Berlin, Heidelberg: Springer Berlin Heidelberg.

Elzinga, C. & Liefbroer, A., 2007. De-standardization of family-life trajectories of young adults: A cross-national comparison using sequence analysis. *… Journal of Population/Revue européenne de …*, 23(3), pp.225–250.

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P., 1996. From data mining to knowledge discovery in databases. *AI magazine*, 17(3), pp.37–54.

Gabadinho, A. et al., 2010. Indice de complexité pour le tri et la comparaison de séquences catégorielles. *Revue des nouvelles technologies de l'information RNTI E*, 19, pp.61–66.

Gabadinho, A. & Ritschard, G., 2009. Mining sequence data in R with the TraMineR package: A users guide for version 1.2. *Department of Econometrics and Laboratory of Demography, University of Geneva, Geneva*.

De Groeve, J. et al., 2015. Extracting Spatio-Temporal Patterns In Animal Trajectories: An Ecological Application Of Sequence Analysis Methods (SAM). *Methods in Ecology and Evolution*, 7(3), pp.369–379.

Halkidi, M., Batistakis, Y. & Vazirgiannis, M., 2001. On Clustering Validation Techniques. *Journal of Intelligent Information Systems*, 17(2-3), pp.107–145.

Hawelka, B. et al., 2015. Collective Prediction of Individual Mobility Traces with Exponential Weights. *submitted*.

Imfeld, S., Haller, R. & Laube, P., 2006. Positional Accuracy of Biological Research Data in GIS–A Case Study in the Swiss National Park. In *7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*. pp. 275–280.

Jain, A.K., Murty, M.N. & Flynn, P.J., 1999. Data clustering: a review. *ACM Computing Surveys*, 31(3), pp.264–323.

Jarboe, G.R. & McDaniel, C.D., 1987. A profile of browsers in regional shopping malls. *Journal of the Academy of Marketing Science*, 15(1), pp.46–53.

Jiang, Q. et al., 2012. Using Sequence Analysis to Classify Web Usage Patterns across Websites. In *2012 45th Hawaii International Conference on System Sciences*. IEEE, pp. 3600–3609.

Kang, H.-Y., Kim, J.-S. & Li, K.-J., 2009. Similarity measures for trajectory of moving objects in cellular space. In *Proceedings of the 2009 ACM symposium on Applied Computing - SAC '09*. New York, New York, USA: ACM Press, p. 1325.

Kaufman, L. & Rousseeuw, P.J., 2009. *Finding Groups in Data: An Introduction to Cluster Analysis*,

Kohonen, T., Kaski, S. & Lappalainen, H., 1997. Self-Organized Formation of Various Invariant-Feature Filters in the Adaptive-Subspace SOM. *Neural Computation*, 9(6), pp.1321–1344.

Kuflik, T. & Dim, E., 2013. Early detection of pairs of visitors by using a museum triage. *Proceedings of the Annual Conference of Museums and the Web*.

Laaksonen, M., 1993. Retail patronage dynamics: Learning about daily shopping behavior in contexts of changing retail structures. *Journal of Business Research*, 28(1-2), pp.3–174.

Laube, P., 2014. *Computational Movement Analysis*, Cham: Springer International Publishing.

Laube, P. & Purves, R.S., 2011. How fast is a cow? Cross-Scale Analysis of Movement Data. *Transactions in GIS*, 15(3), pp.401–418.

Lee Taylor, S. & Cosenza, R.M., 2002. Profiling later aged female teens: mall shopping behavior and clothing choice. *Journal of Consumer Marketing*, 19(5), pp.393–408.

Lee, H.-J. & Joh, C.-H., 2010. Tourism Behaviour in Seoul: An Analysis of Tourism Activity Sequence using Multidimensional Sequence Alignments. *Tourism Geographies*, 12(4), pp.487–504.

Lesnard, L., 2010. Setting Cost in Optimal Matching to Uncover Contemporaneous Socio-Temporal Patterns. *Sociological Methods & Research*, 38(3), pp.389–419.

Levenshtein, V., 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*, 10, pp.707–710.

Levin, S.A., 1992. The Problem of Pattern and Scale in Ecology: The Robert H. MacArthur Award Lecture on JSTOR. *Ecology*, 73(6), pp.1943–1967.

Lin, C.-Y., Peng, W.-C. & Tseng, Y.-C., 2006. Efficient in-network moving object tracking in wireless sensor networks. *IEEE Transactions on Mobile Computing*, 5(8), pp.1044–1056.

Lykourentzou, I. et al., 2013. Improving museum visitors' Quality of Experience through intelligent recommendations: A visiting style-based approach. In *Museums as intelligent environments (MasIE2013)- Workshop co-located with the 9th International Conference on Intelligent Environments - IE'13*.

Manodham, T., Loyola, L. & Miki, T., 2007. A Novel Wireless Positioning System for Seamless Internet Connectivity based on the WLAN Infrastructure. *Wireless Personal Communications*, 44(3), pp.295–309.

Mok, E. & Retscher, G., 2007. Location determination using WiFi fingerprinting versus WiFi trilateration. *Journal of Location Based Services*, 1(2), pp.145–159.

Montello, D.R., 2001. Scale in geography. *International Encyclopedia of the Social and Behavioral Sciences*, pp.13501–13504.

Mount, D.W., 2004. *Bioinformatics : sequence and genome analysis*, Cold Spring Harbor, NY : Cold Spring Harbor Laboratory Press.

du Mouza, C. & Rigaux, P., 2005. Mobility Patterns. *GeoInformatica*, 9(4), pp.297–319.

Mulligan, G.F., 1987. Consumer Travel Behavior: Extensions of a Multipurpose Shopping Model. *Geographical Analysis*, 19(4), pp.364–375.

Nanni, M. & Pedreschi, D., 2006. Time-focused clustering of trajectories of moving objects. *Journal of Intelligent Information Systems*, 27(3), pp.267–289.

Needleman, S. & Wunsch, C., 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48, pp.443–453.

Pelekis, N., Andrienko, G. & Andrienko, N., 2012. Visually exploring movement data via similarity-based analysis. *Journal of Intelligent Information Systems*, 38(2), pp.343–391.

Pihur, V. & Datta, S., 2007. Weighted rank aggregation of cluster validation measures: a monte carlo cross-entropy approach. *Bioinformatics*, 23(13), pp.1607–1615.

Poole, M.S. & Holmes, M.E., 1995. Decision Development in Computer-Assisted Group Decision Making. *Human Communication Research*, 22(1), pp.90–127.

Prinzie, A. & Van den Poel, D., 2006. Incorporating sequential information into traditional classification

models by using an element/position-sensitive SAM. *Decision Support Systems*, 42(2), pp.508–526.

Rand, W.M., 1971. Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, 66(336), pp.846–850.

Rekimoto, J., Miyaki, T. & Ishizawa, T., 2007. LifeTag: WiFi-based continuous location logging for life pattern analysis. In *Location- and Context-Awareness third international symposium, LoCA*. pp. 35–49.

Ren, Y. et al., 2015. Analyzing Web Behavior in Indoor Retail Spaces. *Journal of the Association for Information Science and Technology*.

Ren, Y., Tomko, M. & Ong, K., 2014. The influence of indoor spatial context on user information behaviours. *Workshop on Information Access in Smart Cities, held at ECIR*.

Ricci, F. et al., 2011. *Recommender Systems Handbook*, Boston, MA: Springer US.

Richter, K.-F., Schmid, F. & Laube, P., 2012. Semantic trajectory compression: Representing urban movement in a nutshell. *Journal of Spatial Information Science*, 2012(4), pp.3–30.

Rousseeuw, P., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20, pp.53–65.

Rykiel, E.J., 1996. Testing ecological models: the meaning of validation. *Ecological Modelling*, 90(3), pp.229–244.

Sapiezynski, P. et al., 2015. Tracking Human Mobility Using WiFi Signals. *PLoS ONE*, 10(7).

Sheth, J. & Park, C., 1974. A theory of multidimensional brand loyalty. In *Association for Consumer Research*. pp. 449–459.

Sheth, J. & Raju, P., 1974. Sequential and cyclical nautre of information processing models in repetitive choice behavior. *Advances in Consumer Research*, 1(1), pp.348–358.

Shilton, K., 2009. Four billion little brothers? *Communications of the ACM*, 52(11), p.48.

Shoval, N. et al., 2015. The application of a sequence alignment method to the creation of typologies of tourist activity in time and space. *Environment and Planning B: Planning and Design*, 42(1), pp.76–94.

Shoval, N. & Isaacson, M., 2007. Sequence alignment as a method for human activity analysis in space and time. *Annals of the Association of American Geographers*, 97(2), pp.282–297.

Silberschatz, A. & Tuzhilin, A., 1996. What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge and Data Engineering*, 8(6), pp.970–974.

Spaccapietra, S. et al., 2008. A conceptual view on trajectories. *Data & Knowledge Engineering*, 65(1), pp.126–146.

Stafford, T.F., Turan, A. & Raisinghani, M.S., 2014. International and Cross-Cultural Influences on Online Shopping Behavior. *Journal of Global Information Technology Management*, 7(2), pp.70–87.

Stovel, K. & Bolan, M., 2004. Residential Trajectories: Using Optimal Alignment to Reveal The Structure of Residential Mobility. *Sociological Methods & Research*, 32(4), pp.559–598.

Thompson, J., Higgins, D. & Gibson, T., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix. *Nucleic acids research*, 22, pp.4673–4680.

Tomko, M. et al., 2014. *Large-Scale Indoor Movement Analysis: The data, context and analytical challenges*,

Wesley, S., LeHew, M. & Woodside, A.G., 2006. Consumer decision-making styles and mall shopping behavior: Building theory using exploratory data analysis and the comparative method. *Journal of Business Research*, 59(5), pp.535–548.

Wilson, C., 2008. Activity patterns in space and time: calculating representative Hagerstrand trajectories. *Transportation*, 35(4), pp.485–499.

Wilson, C., 2001. Activity patterns of Canadian women: Application of ClustalG sequence alignment software. *Transportation Research Record: Journal of the Transportation Research Board*, 1777, pp.55–67.

Wilson, C., 2006. Reliability of sequence-alignment analysis of social processes: Monte Carlo tests of ClustalG software. *Environment and Planning A*, 38(1), pp.187–204.

Wilson, C., Harvey, A. & Thompson, J., 1999. ClustalG: Software for analysis of activities and sequential events. In *Workshop on Longitudinal Research in Social Sciences*.

Wilson, W., 1998. Activity pattern analysis by means of sequence-alignment methods. *Environment and Planning A*, 30(6), pp.1017–1038.

Woo, S. et al., 2011. Application of WiFi-based indoor positioning system for labor tracking at construction sites: A case study in Guangzhou MTR. *Automation in Construction*, 20(1), pp.3–13.

Yeung, K.Y., Fraley, C., et al., 2001. Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17(10), pp.977–987.

Yeung, K.Y., Haynor, D.R. & Ruzzo, W.L., 2001. Validating clustering for gene expression data. *Bioinformatics*, 17(4), pp.309–318.

Yuan, Y. & Raubal, M., 2014. Measuring similarity of mobile phone user trajectories–a Spatio-temporal Edit Distance method. *International Journal of Geographical Information Science*, 28(3), pp.496–520.

Zafarani, R. & Liu, H., 2015. Evaluation without ground truth in social media research. *Communications of the ACM*.

Zandbergen, P.A., 2009. Accuracy of iPhone Locations: A Comparison of Assisted GPS, WiFi and Cellular Positioning. *Transactions in GIS*, 13, pp.5–25.

Zeithaml, V.A., 1985. The New Demographics and Market Fragmentation. *Journal of Marketing*, 49(3), pp.64–75.

References

# PERSONAL DECLARATION

I hereby declare that the submitted thesis is the result of my own, independent work. All external sources are explicitly acknowledged in the thesis.

June 30, 2016

Simon Jakob