Uncovering movement suspension patterns on GPS tracking data.

*Analysing the temporal dimension of movement suspension patterns on pedestrian movement*

Michail Marinakis

March 2011
Uncovering movement suspension patterns on GPS tracking data.

Analysing the temporal dimension of movement suspension patterns on pedestrian movement

Michail Marinakis

Registration Number 830215543040

Supervisors:

Arend Ligtenberg
Daniel Orellana

A thesis submitted in partial fulfilment of the degree of Master of Science at Wageningen University and Research Centre, The Netherlands

April 2011
Wageningen, the Netherlands

Thesis code number: GRS-80436
Wageningen University and Research Centre
Laboratory of Geo-Information Science and Remote Sensing
Table of Contents

LIST OF FIGURES .............................................................................................................VI
LIST OF TABLES ................................................................................................................VI
ACKNOWLEDGEMENTS ..................................................................................................VI
ABSTRACT .........................................................................................................................VIII

1 INTRODUCTION ...........................................................................................................1
   1.1 BACKGROUND ......................................................................................................1
   1.2 PROBLEM DEFINITION .....................................................................................1
   1.3 RESEARCH OBJECTIVE AND RESEARCH QUESTIONS ......................................4

2 METHODOLOGY .........................................................................................................4
   2.1 INTRODUCTION ..................................................................................................4
   2.2 DETECTING MOVEMENT SUSPENSION ..............................................................6
   2.3 CALCULATION OF CLUSTERS .........................................................................6
       2.3.1 Introduction .................................................................................................6
       2.3.2 KDE+AHC Variation ..................................................................................7
       2.3.3 AHC Variation ............................................................................................9
       2.3.4 Evaluation of Clustering ............................................................................10
   2.4 TEMPORAL PATTERNS .....................................................................................10
   2.5 ASSIGNING GEOGRAPHICAL ELEMENTS TO STOPS ......................................10
   2.6 ASSIGNING ACTIVITIES TO STOPS .................................................................11

3 RESULTS AND DISCUSSION .....................................................................................11
   3.1 STUDY AREA: DWINGELDERVELD NATIONAL PARK .......................................11
   3.2 DATA AND SOFTWARE ....................................................................................11
       3.2.1 Experiment 2006 ......................................................................................11
       3.2.2 Experiment 2010 .....................................................................................12
   3.3 EVALUATION OF METHODOLOGY ................................................................12
       3.3.1 Control Dataset .......................................................................................12
       3.3.2 Test dataset ...............................................................................................12
       3.3.3 Comparing “KDE+AHC” and “AHC” variation .......................................13
       3.3.4 Defining the number of clusters ..............................................................15
   3.4 TEMPORAL PATTERNS ....................................................................................18
       3.4.1 Introduction ...............................................................................................18
       3.4.2 Distribution of stops in different periods .................................................19
   3.5 GEOGRAPHICAL ELEMENTS (DESTINATIONS, CONNECTIVITY OF PATHS) ........24
   3.6 ACTIVITIES .......................................................................................................28

4 CONCLUSION AND RECOMMENDATIONS ..........................................................30
   4.1 CONCLUSIONS ..................................................................................................30
   4.2 RECOMMENDATIONS ........................................................................................32

5 REFERENCES .............................................................................................................32
List of Figures

Figure 1.1 Space-Time Path (Vrotsou, Ellegård, & M., 2007) ................................................................. 2
Figure 2.1 Methodology Flowchart ........................................................................................................ 5
Figure 2.2 Percentage Volume Contour of 95% of the vectors, calculated from Kernel density estimator ................................................................................................................................. 7
Figure 2.3 Agglomerative hierarchical clustering on data objects {a, b, c, d, e} (Han & Kamber, 2006). 8
Figure 2.4 Dendrogram graph. 14 clusters could be recognized at the level of the red line. ............. 8
Figure 2.5 Number of clusters for each clustering step of AHC algorithm ........................................... 9
Figure 2.6 Percentage of valid points inside buffers with different distances ....................................... 14
Figure 2.7 Spatio-temporal clusters with specific number of persons within them ............................... 16
Figure 3.1 Test 543 and Test 159. Example of mean centres of individual stops and mean centre of the set of pictures .................................................................................................................................................. 16
Figure 3.2 Frequency distribution of the number of stops per visitor ................................................. 17
Figure 3.3 Total number of stops in different time periods ................................................................. 19
Figure 3.4 Average number of stops per visitor .................................................................................... 19
Figure 3.5 Average percent of stopping time (% of total time staying in the park) .............................. 21
Figure 3.6 Number of stops with encounter (% of total stops) .......................................................... 21
Figure 3.7 Number of stops per hour .................................................................................................... 22
Figure 3.8 Number of visitors per hour ............................................................................................... 22
Figure 3.9 Average duration of stops and encounters (in minutes) .................................................... 23
Figure 3.10 Individual stops in which a encounter occurred (% of total number of stops) ...................... 23
Figure 3.11 Individual stops with encounter (% total number of stops) that occurred at each specific sequence of stops ........................................................................................................................................ 24
Figure 3.12 Specific amenities where stops occurred with more than two park features included ...... 25
Figure 3.13 Number of individual stops per park feature ................................................................. 26
Figure 3.14 Average duration of individual stops per park feature .................................................... 26
Figure 3.15 Radio-telescope, Visitor Centre and Sheep Farm ............................................................. 27
Figure 3.16 Number of visitors per activity ......................................................................................... 28
Figure 3.17 Average duration of non-stopping time and stopping time .............................................. 29
Figure 3.18 Average number of individual stops per person ............................................................... 30

List of Tables

Table 3.1 Two level of dissimilarities after performing KDE+AHC variation ...................................... 13
Table 3.2 Two level of dissimilarities after performing AHC variation .............................................. 13
Table 3.3 Comparison of the two methodology variation ..................................................................... 15
Acknowledgements

This thesis is the result of a lot of working hours trying to create something new and correcting my own mistakes, a process that taught me many things. I would like to thank my supervisors, Arend for his critical advice and guidance and Daniel for his ideas, support, unlimited patience and lot of hours of his time in order to help me writing this report. Also I am thankful to my friends and fellow students who helped me during this stressful but creative period.
Abstract

The study of pedestrians’ movement is an important subject in transportation analysis. There is a lack of integration of temporal information in the study of spatial phenomena related with movement. The technological advancement in tracking technology and in computer science assists to collect and analyse spatio-temporal data of moving entities. This report formulates a space-time approach that could be used for exploratory analysis of spatio-temporal activity data. Our scope is to analyse the temporal dimension of movement suspension patterns (MSP). MSP are spatial clusters of low speed vectors associated to the collective stopping behaviour of pedestrians (Orellana & Wachowicz, in press).

In this research, we analyse the MSP in two movement datasets collected in the Dwingelderveld National Park. We use Agglomerative Hierarchical Clustering (AHC) to calculate spatio-temporal clusters on those MSP and finally extract individual stops. Furthermore, these stops were associated with different geographical elements in the park (e.g. park features, cross paths, path connectivity) and pedestrians’ activities.

We concluded that the number of visitors’ stops, their duration, and the number of encounters between the visitors that occurred, were affected by the different attractions and facilities in the park but also by the activities that these visitors had. The combination of LISA with AHC is a consistent approach for exploring pedestrian’s stopping behaviour in space and time. This approach could be used to optimize the design of location-based-services but also could be applied as a tool in different scientific fields such as traffic management, animal behaviour and urban planning.
1 Introduction

1.1 Background

In recent years Global Positioning System (GPS) receivers have become smaller, cheaper and increasingly integrated into a number of consumer products: on-board navigation, small computers, cell phones, etc (Nielsen & Hovgesen, 2004). These tools have been used not only for navigation purposes but also for collecting spatial data and information about user’s locations and itineraries. The technological advancement in tracking technology reached a level that allows the seamless tracking of individuals needed for the analysis of movement patterns (Gudmundsson, Laube, & Wolle, 2008). The definition of movement patterns varies through the literature. One definition of movement pattern is: “any high-level description of the movement of an individual or a group of individuals. This description can but must not relate the movement to the underlying space” (Laube, 2009).

Hägerstrand’s argument for taking time into account along with space was that “time has a critical importance when it comes to fitting people and things together for functioning in socio-economic systems, whether these undergo long-term changes, or rest in something which could be defined as a steady state.” (Hägerstrand, 1970). Our lives consist of activities such as working, socializing, shopping, and recreation that require resources that are available only at a few locations and for limited durations (Miller, 2005b). Walking is one of these activities. It can be undertaken for recreation (i.e. for leisure or exercise) or transport e.g. to the shops, schools or to public transportation (Bentley, Jolley, & Kavanagh, 2010). At pedestrian movement, stops occurred in order to carry out different kind of activities. These stops could also be referred as movement suspension (Orellana & Wachowicz, in press). The purpose of this research is to analyse pedestrians’ suspension of movement during walking for recreation in order to comprehend their interaction with environmental factors. The integration of temporal data into the analysis of spatial information will contribute to a further understanding of pedestrian’s spatial behaviour.

This report is organized as follows. In the next subsection, we present the problem definition in the context of a review of the research related to space-time geo-information systems and spatio-temporal patterns analysis. The objective and the main goals of this report are stated in subsection 1.3. In chapter 2 we present the methodology that we used. In chapter 3 we present the evaluation of our methodology, the implementation of it and the results. Concluding remarks are presented in the final section.

1.2 Problem Definition

“Historically, social scientists studying the effects of space on human behaviour tended to treat time as an external factor, something that is relevant to understanding a given phenomenon, but not essential” (Corbett, 2001). At 1970, Time Geography was introduced by Hägerstrand. He was supporting that we need to understand better what it means for a location to have not only space coordinates but also time coordinates. In his framework,
time is included as a third vertical axis added to a two dimensional system which represents space. In the two dimensional system we can visualize and measure the location of the object and its movement. At the vertical axis we can represent it as a progressively happening in time. This information can be related to different human activities which can be visualized by the space-time path.

Figure 1.1 Space-Time Path (Vrotsou, Ellegård, & M., 2007)

A space-time path is the container of all activities performed by a person, since all activities take place at certain locations and time periods. Each of these activities occupies a portion of the space-time path. (Yu, 2006). In Figure 1.1 a space-time path of a single person is presented. The parallel lines to the time axis are stops of this person at different locations e.g. home, work, shop. Time geography also suggests that individual’s activities in space and time are conditioned by three types of constraints – capability constraints, authority constraints, and coupling constraints (Hägerstrand, 1970).

“The extension of a two-dimensional map with a third orthogonal time axis produces a very powerful tool for uncovering movement patterns” (Gudmundsson, Laube, & Wolle, 2008). Many researchers from different fields have found the time-geographic framework useful to comprehend human activities and travel behaviour (Ellegård, 1999; Kwan, 2004, 2007). They calculate movement patterns which could be used as input to some decision making process (e.g. to derive useful knowledge for optimizing traffic management), to acquire more knowledge about the travelling objects (e.g. analysing bird trajectories), or to control the proper implementation of transportation logistics (e.g. monitoring worldwide delivery of parcels in a courier company) (Spaccapietra et al., 2008). “Research frontiers implied by the time-geographic measurement framework include query design, mapping the theory to networks, extending the theory to velocity fields, imperfect measurement, and incorporating virtual interaction” (Miller, 2005a). Spatial data mining, in which query design is involved, focuses on searching rules of the geographical statement, the structures of distribution and the spatial patterns of phenomena. However, many methods ignore the temporal information; thus, limited results are describing the statement of spatial phenomena (Chai, Su, & Ma, 2010).

An approach for analysing movement data is to treat them as trajectories. We can see a trajectory as a sequence of moves going from one stop to the next one (or as a sequence of stops separating the moves) (Spaccapietra et al., 2008). The visualization of the data in a
space-time cube has been used also as a tool for analysing movement patterns and through that, activities of moving objects (Gatalsky, Andrienko, & Andrienko, 2004). The classical space-time paths track individuals over time. One approach of including groups of individuals is the “generalized space-time path”. A generalized space-time path shows the changes in spatial distribution patterns of distinct subgroups of individuals between different time periods (Shaw, Yu, & Bombom, 2008).

One more approach to analyse collective movement is the “movement suspension patterns” (MSP) (Orellana and Wachowicz, in press). In this research the MSP was used to detect the location of stops for a group of people visiting a natural area. In their approach the GPS recordings were treated as movement vectors. The movement vectors have a specific direction starting from each GPS point and formulating a line segment in the Euclidean space. They comprise a recorded space-time location and computed movement parameters, which are displacement (absolute distance from previous point), time step (absolute time from previous point), speed (displacement over time), and bearing (angle between two consecutive points with respect to true North) (Orellana & Wachowicz, in press). Stops are conceptualised as movement suspension, represented by clusters of low speed movement vectors with high spatial correlation. In their work, the authors only consider the spatial properties of the movement dataset, leaving the temporal dimension unexplored. In our research, the time is included in the analysis to improve the interpretation of movement suspension patterns. Knowing the direction and the speed of an observation from the GPS we can use movement vectors in order to define the spatial and temporal correlation not only of individuals but also of groups of people. According to Dodge et al, spatio-temporal data have specific parameters which are speed and velocity (i.e. rate of change of position and direction) and these parameters can be derived from changes in both spatial position and time instances (Dodge, Weibel, & Lautenschutz, 2008).

Furthermore, clustering process is necessary to analyse and explore large amount of movement data, considering also groups of objects and not only individuals. Clustering algorithms that have been widely used for analysing movement data are based on DBSCAN (Density-Based Spatial Clustering of Applications with Noise). The key idea is that for each point of a cluster the neighbourhood of a given radius has to contain at least a minimum number of points (Ester, Kriegel, Sander, & Xu, 1996). However DBSCAN cannot discover clusters with large variance in density because it depends on the initial value for its parameters, radius and minimum number of points (Ram, Jalal, Jalal, & Kumar, 2010). ST-DBSCAN has been developed for discovering clusters according to non-spatial, spatial and temporal values of the objects (Birant & Kut, 2007). Also OPTICS generates a data structure that allows one to calculate efficiently the result of DBSCAN for any desired density threshold (Stuetzle, 2003). It has also been used in combination with interactive visual displays (Andrienko & Andrienko, 2009). All of these methods need from the user to develop a threshold in distance, time or minimum number of neighbours at the beginning of the calculations. In other cases, methods of clustering such as k-means need to know the number of clusters because is used as an input at the algorithm (Shoshany, Even-Paz, & Behjor, 2007). Clustering methods are essential for revealing movement patterns. The overall challenge consists on relating these movement patterns with the underlying
geography, in order to understand where, when and ultimately why the entities move the way they do (Gudmundsson, Laube, & Wolle, 2008).

1.3 Research Objective and research questions

The research objective is:

“To explore how clustering techniques facilitate the analysis of the temporal dimension of movement suspension patterns and their relation with environmental features and pedestrians’ activities”

This objective will be covered by answering the following research questions:

a) How the movement suspension patterns are formed in space and time?

b) Are there geographical elements (e.g. functional features, commercial and common facilities) related to the spatio temporal patterns of movement suspension?

c) Is it possible to associate the detected movement suspension patterns to specific pedestrian activities?

2 Methodology

2.1 Introduction

It is necessary to define some terms that will be used often in the next chapters:

a) Movement Suspension Patterns: In a movement dataset, Movement Suspension Patterns refer to the vectors classified as suspension, having a speed below the mean and a statistically significant spatial association (Orellana & Wachowicz, in press). In this research, the term "movement suspension" indicates the stopping behaviour of pedestrians. (Orellana & Wachowicz, in press).

b) Spatial and spatio-temporal clusters: Vectors classified as movement suspension patterns can be clustered in space or space and time. In this research two variations of a clustering method are proposed (described below). Clusters may include movement vectors of one or more visitors.

c) Individual stops: When the clusters are associated to each unique visitor, they represent individual stops. In this report, we use the term stops for the sake of simplicity.

A flowchart which represents an overview of our methodology is depicted in Figure 2.1. Firstly, we clustered movement suspensions patterns, as these were formulated from LISA process. We used two different clustering techniques, KDE+AHC and AHC to define visitors’ stops. Secondly, we evaluated the results of these two techniques and we decided to continue with the implementation of AHC variation in our study area. Our final results are related with the temporal distribution of the visitors’ stop, the geographical elements where the stops occurred and the pedestrians’ activities during the stops.
Figure 2.1 Methodology Flowchart
2.2 Detecting movement suspension

The method suggested by Orellana and Wachowicz (in press) is selected to detect the locations where a group of people suspended their movement. This method is based on Exploratory Spatial Data Analysis approaches, combined with the use of Local Indicators of Spatial Association (LISA), both proposed by Anselin at 1993 and 1995 respectively. Exploratory Spatial Data Analysis can be considered as data-driven analysis, which is applied without including many preconceived ideas, theories or hypotheses. It aims to describe spatial distribution, discover patterns of spatial association and outliers (Anselin, 1993). LISA is any statistic that satisfies the following requirements:

a. The LISA of each observation gives an indication of the extent of significant spatial clustering of similar values around that observation;
b. The sum of LISAs for all observations is proportional to a global indicator of spatial association (Anselin, 1995).

In this research, the LISA is based on the Moran’s I global summation of individual statistics to evaluate the spatial association by calculating Local Moran’s I and evaluating the statistical significance of each unit. The outputs of this process are three values which are computed for each movement vector: the Z score values, the P-values the LISA values. The Z score values and the p-values are a test of statistical significance that helps to decide whether or not to reject the null hypothesis. The null hypothesis for pattern analysis essentially states that there is no spatial pattern among the features, or among the values associated with the features, in the study area (Mitchell, 2005). The P-values are the probabilities and the Z values are measures of standard deviation. Both statistics are associated with the standard normal distribution. Therefore, the value of LISA indicates local association of speed values. For example, high positive values imply that a movement vector is surrounded by vectors with similar values; meanwhile high negative values indicate that a movement vector is surrounded by very different values (Orellana & Wachowicz, in press). Moreover, the Z value gives an indication of the statistical significance of the computed value and can be used to select the data corresponding with a certain confidence level (e.g. 5% corresponds to a Z value of standard deviation 1.96). In this case, movement suspension patterns should be spatial clusters of low-speed vectors with high statistical scores of spatial association (Orellana & Wachowicz, in press).

2.3 Calculation of Clusters

2.3.1 Introduction

The output of the approach by Orellana & Wachowicz is a set of movement vectors with a new attribute stating for each vector if it is classified as movement suspension or not. When vectors classified as suspension, are plotted in a geographical space and form spatial clusters. However, the spatial and temporal extension and boundaries of those clusters are not explicitly defined. A proper definition of the spatial clusters will improve the exploratory analysis and ultimately help to associate them with geographical features and pedestrian activities.
We propose a method with two variations in order to define the extension and boundaries of the movement suspension clusters. The first variation ("KDE+AHC") combines Kernel Density Estimator (KDE) and Agglomerative Hierarchical Clustering (AHC). The second variation ("AHC") uses only AHC to define the clusters. KDE is one of the most popular methods for analysing the underlying properties of point events and measures the variation in the mean value of the process (Silverman, 1986, Xie & Yan, 2008,). Also the implementation of this estimator is easy and fast. AHC is a clustering method in which, unlike the partitional clustering methods, the number of clusters is not an input for the algorithm. In addition, the algorithm can get arbitrary shapes of clusters and it generates an ordering of the objects, which may be useful for data interpretation and display.

2.3.2 KDE+AHC Variation

The KDE+AHC variation consists of two steps. The first step defines the spatial component of the clusters and the second defines the temporal component. In the first step spatial clusters are formulated as hot spots of movement suspension patterns (Figure 2.2) and associated with the spatial characteristics of the area. These clusters are defined by using Kernel Density Estimation (KDE) and Percent Volume Contours (PVC). The KDE is based on the quadratic kernel function or also called Epanechnikov Kernel (Silverman, 1986). A Percent Volume Contour represents the boundary of the area that contains x% of the volume of a probability density distribution (Beyer, 2004). We used the 95% volume contour which contains 95% of the points that were used to generate the KDE, so small clusters with few observations and individual vectors outside the kernel of the clusters are not included.

Figure 2.2 Percentage Volume Contour of 95% of the vectors, calculated from Kernel density estimator

In the second step, each spatial cluster is divided into smaller clusters using Agglomerative Hierarchical Clustering (AHC) with the method of Single Linkage, also known as Nearest Neighbour. For this step, only the temporal dimension is analysed.

The algorithm of AHC constructs a hierarchy of clusters. At the beginning, each point is taken as an individual cluster and gradually merged with each other to form new clusters until at
the end they represent only one cluster which includes all the points (Figure 2.3). At each step the two more similar clusters are merged to form one larger cluster (Kaufman & Rouseeuw, 1990). We used the Nearest Neighbour method computing the Euclidean distance in the temporal dimension. In the Nearest Neighbour method, the distance between two clusters is the minimum of the distances between all pairs of patterns drawn from the two clusters (one pattern from the first cluster, the other from the second) (Jain, Murty, & Flynn, 1999). Although it suffers from a chaining effect (Nagi, 1968), the AHC algorithm is versatile and it is widely used in different scientific fields such as crime analysis (Zeng & Chen, 2004) and epidemiology (Zeng D., Chen, Lynch, Eidson, & Gotham, 2005).

![Figure 2.3 Agglomerative hierarchical clustering on data objects {a, b, c, d, e} (Han & Kamber, 2006).](image)

The implementation of the algorithm provides a dendrogram graph representing the nested grouping of patterns and dissimilarity levels at which grouping change. For example, in Figure 2.4 a dendrogram of 37 vectors is presented. At point zero of the y-axis of the graph, each vector represents a unique cluster.

![Figure 2.4 Dendrogram graph. 14 clusters could be recognized at the level of the red line.](image)

At the first level of clustering (red line) 14 clusters are formulated and the graph continues until only one cluster is represented from these points. This graph can be cut at a desired dissimilarity level forming a number of clusters identified by simply connected components (Jain, Murty, & Flynn, 1999). The lowest height at this example is at 0.12 and the largest is at 1.38 where only one clusters is formed above that height. The height of the node of two or more elements can be considered as proportional to the Euclidean distance value between
two or more components that are clustered together at that level. This distance is calculated from the Nearest Neighbour method. Moreover, the height on the dendrogram does not have any units and it is mainly for representative purposes. We assume that the optimal cluster configuration can be recognized only by subjective interpretation and highly depend on the application (Jung, Park, & Du, 2003). In Figure 2.5 is depicted the number of clusters in each clustering step of the dendrogram of Figure 2.4.

![Figure 2.5 Number of clusters for each clustering step of AHC algorithm](image)

For analysis purposes, the movement suspension for each visitor (i.e. individual stops) must be defined. Therefore, each cluster is divided according each visitor. For example if the identification number of the cluster is 100 and from that cluster, visitors with the ID 33 and 35 stopped by, then two new clusters are formulated: 33_100 and 35_100. Each resulting cluster is therefore considered as an individual stop. Afterwards, we computed the duration of each stop by subtracting the first from the last time of the vectors belonging to it. To define if in that spatio-temporal cluster, two or more visitors had an encounter, we calculated the total duration of that cluster. Then we sum all the duration of individual stops that belonged to this specific spatio-temporal cluster and we compared it with the total duration of the spatio-temporal cluster. If the sum was larger than the duration of the spatio-temporal cluster then it means that there was some overlap between these individual stops so there was an encounter for at least two visitors in that spatio-temporal cluster. In addition, some other characteristics were defined, such as the sequence of the stops per visitor and the number of vectors in each individual stop.

### 2.3.3 AHC Variation

The “AHC” variation consists on applying the Agglomerative Hierarchical Clustering (AHC) to the spatial and temporal components of the vectors simultaneously. This means that the coordinates of each vector in space \((x, y)\) and time \((t, t)\) are used to compute the Euclidean distance for the similarity function. It is necessary to consider the equivalence of the input dimensions for the clustering process. Since we dealt with spatial and temporal dimension that are not equivalent, we had to standardize them. The standardization consists on
subtracting the variable's mean value and dividing by the variable's mean absolute deviation. Finally, the individual stops were computed as explained above.

2.3.4 Evaluation of Clustering

The data from a controlled experiment was used to evaluate the results of our method. In this experiment, a group of pedestrians walked and stopped in designated locations. The evaluation consists on comparing the test dataset with a control dataset. The test dataset is composed by the mean centres of the individual stops computed with the two variations presented above. The control dataset is composed by a set of recorded locations were the participants actually stopped.

We took into account the vectors that were calculated as movement suspension using the LISA process and the clusters that were formed after the implementation of the KDE+AHC and AHC variation. After the separation of spatio-temporal clusters to individual stops, we calculated the mean centre of these individual stops. The evaluation consists on computing the distance between the mean centres of individual stops with the locations of the corresponding control stops. If this distance was below a certain threshold, the result was classified as true positive. We performed this process in two different levels of dissimilarity as these levels were defined by the dendrogram of the two variations.

2.4 Temporal Patterns

For deriving temporal patterns, data mining concepts are helpful to reveal the information we want. We can derive differences in several variables of stops by comparing different periods of time. These variables were associated with the number of stops in total and per visitor, the duration of the stops, and frequency of stops in which co-occurrence in space and time between two or more visitors occurred (encounters). The different time periods were months (May and August), weekends and weekdays, and hours of day. Specifically, some hypothesis should be made, for example, take the time period as a day and the areas of interest as the information points of the national park. Patterns could formulate and this time period can give us interesting results. If not, then another time period can be formulated.

2.5 Assigning geographical elements to stops

Pikora et al (2003) developed a framework that identifies four environmental components which are related with people’s walking: functional (physical attributes of the street and path that reflect the structural aspects of the environment), safety (characteristics of areas that provide safe physical environments e.g. lighting and traffic safety such as crossings), aesthetics (presence, condition and size of trees, parks, gardens, levels of pollution but also the diversity of natural sights and architectural designs) and destinations (availability of commercial and community facilities) (Pikora, Giles-Corti, Bull, Jamrozik, & Donovan, 2003).

We focused at the functional features of the environment and the destinations of pedestrians either final or intermediate, since our research is more spatially-oriented and these factors could be derived from GIS available data. The functional features include
locations of cross-paths and land use. The destinations refer to different park features. A buffer was created for each individual stop and spatially compared with the set of park features. Besides the park features, we created two more categories to assign to the individual stops; one is the “Undefined” environmental feature where non-obvious outcome can be assigned to that stop after the previous process and the “Unknown” environmental feature where there was not any park feature or any path intersection to assign for that stop.

2.6 Assigning activities to stops

A survey was conducted during the experiment in 2006. This questionnaire included questions about motivations, environmental values, behaviour (e.g. entry point, destinations, attractions visited, main activities), special places, and socio-demographics (van Marwijk, 2009). We focused at the activities of pedestrians. These activities were divided in 11 main categories. The categories are: walking/hiking, picnicking, observe flora/fauna, dog walking, sunbathing/relaxing, sports (running), taking pictures, visiting restaurant, going to visitor centre, visiting sheep farm and other (open answer). We associated these activities both with the environmental factors and with the characteristics of the spatio-temporal patterns which are described above. This combination gives us a comprehensive picture about the spatial behaviour of the pedestrians in the park.

3 Results and Discussion

3.1 Study Area: Dwingelderveld National Park

Dwingelderveld National Park is in the north east of Netherlands in the provenance of Drenthe and was established in 1991. This park covers an area of 3,700 hectares and contains the largest wet heathland in Western Europe (Samenwerkingsverband Nationale Parken, 2005). There are more than 60 bog pools on the heaths and in the woods but also some juniper shrubs. Dwingelderveld attracts over 1.5 million visitors each year. Sixty kilometres of walking trails and forty kilometres of cycle paths allow visitors to explore the area (Samenwerkingsverband Nationale Parken, 2005). The park contains different kind of amenities for the visitors. These amenities are wetlands, sheep farms, bird-watching lookouts, information centres, a tea house, a snack bar and some cultural spots such as a historical house and a radio-telescope (van Marwijk, 2009). There are several access points to the park, many of them near parking facilities.

3.2 Data and Software

3.2.1 Experiment 2006

The data of the experiment 2006 comes from different datasets. The first dataset is a GPS tracking dataset collected during spring and summer of 2006 for seven days including weekdays and weekends (van Marwijk, 2009). This data refers to 372 visitors and contain 141,824 GPS recordings. The second dataset describes the visitors’ activities during the time that they spent in the park, based on a survey applied to the visitors who carried the GPS
devices (van Marwijk, 2009). The third dataset consists on the locations and descriptions of several park features collected from several specialised web pages containing recommendations and tips for visiting the Dwingelderveld National Park (Orellana & Wachowicz, in press). The last dataset contains the structure and the condition of the path network of the national park. We used this data for deriving association between geographical elements and pedestrians’ activities with individual stops.

3.2.2 Experiment 2010

The data for the validation of our two methodology variations was collected during a controlled experiment at December 12, 2010 at the study area. It consists from two different datasets, the test dataset and the control dataset. The test dataset contains 25.138 GPS recordings collected during the experiment with 28 persons. The participants were walking in couples at predefined routes inside the Dwingelderveld National Park. Each couple was instructed to stop for one minute in predefined places, to mark that position at the GPS, and to take 4 pictures with a camera (one of the GPS screen and 3 of the place). After that they had to walk to the next stop location. We used this dataset to implement our methodology variations and define individual stops. The control dataset consists from the positions of where the pictures were taken. We used this dataset to calculate for every set of pictures that were taken per person the mean centre of them. These mean centres represent the location of the corresponding control stops.

All the data were processed and stored using ArcGIS system. For the implementation of the clustering process and for statistical computation we use R as programming language and software environment.

3.3 Evaluation of methodology

3.3.1 Control Dataset

For the evaluation we used the two datasets from experiment 2010. The control dataset indicated the mean centres of the four pictures taken in each predefined position. This dataset was composed by 456 positions. Each mean centre represented one stop of each person. There was an average of 15.53 individual stops per person and the average duration of these stops was 2.02 minutes. 51 of them (11.18%) were validated as false negative stops and the rest 405 (88.82%) as true positive after associated with the movement suspension patterns of LISA. The results of the MSP analysis were compared with the time of those pictures, if they corresponded, the result was a True Positive and if a control stop was not detected then it was a False Negative.

3.3.2 Test dataset

The test dataset was used as an input for the two variations KDE+AHC and AHC. We calculated spatio-temporal clusters and then stops, in two different dissimilarity levels for each variation. In Table 3.1 the results after implementing KDE+AHC variation are depicted. The first dataset, KDE+AHC test 392 (named after the numbers of spatio-temporal clusters obtained), was computed according to the lowest level of dissimilarity in the dendrogram of
AHC algorithm. 392 spatio-temporal clusters and 809 individual stops were identified. The second test dataset, *KDE+AHC test 168* was computed with the second level of dissimilarity at the dendrogram. The number of spatio-temporal clusters in that case is 168 and the individual stops 433.

<table>
<thead>
<tr>
<th>Total Number of stops</th>
<th>AVG duration per stop (in minutes)</th>
<th>Stops per person</th>
<th>AVG number of vectors per stop</th>
<th>AVG number of person per spatio-temporal cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>KDE+AHC test 392</td>
<td>809</td>
<td>1.97</td>
<td>28.89</td>
<td>6.64</td>
</tr>
<tr>
<td>KDE+AHC test 168</td>
<td>433</td>
<td>4.33</td>
<td>15.46</td>
<td>12.41</td>
</tr>
</tbody>
</table>

Table 3.1 Two level of dissimilarities after performing KDE+AHC variation

We followed exactly the same process for the AHC variation and the results are presented in Table 3.2. Again the names of the tests were formulated from the number of spatio-temporal clusters that we calculated in each test. In the *AHC test 543*, the number of individual stops was 832 and in the *AHC test 159* the stops were 441. In both variations, KDE+AHC and AHC, the differences between the two levels of dissimilarities were major. In the next chapter we will analyse further how to define the optimal number of spatio-temporal clusters.

<table>
<thead>
<tr>
<th>Total Number of stops</th>
<th>AVG duration per stop (in minutes)</th>
<th>Stops per person</th>
<th>AVG number of vectors per stop</th>
<th>AVG number of person per spatio-temporal cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>AHC test 543</td>
<td>832</td>
<td>1.8</td>
<td>29.7</td>
<td>6.51</td>
</tr>
<tr>
<td>AHC test 159</td>
<td>441</td>
<td>4.25</td>
<td>15.75</td>
<td>12.29</td>
</tr>
</tbody>
</table>

Table 3.2 Two level of dissimilarities after performing AHC variation

3.3.3 Comparing “KDE+AHC” and “AHC” variation

After comparing the two tables above, we could observe that both methods are following the same tendency. In the second level of dissimilarity in both variations, the results were closer with the control dataset than in the first level. Furthermore when we compared only the second level of dissimilarity between the different variations, we could see that the differences were, in absolute values, between 0.018% and 1.88%.

The next step of our comparison was to create buffers with different radii around the mean centres of the individual stops derived from the test dataset. We could then calculate the percentages of control stops that were included within these buffers. In Figure 3.1 the tests of both variations are presented.
Figure 3.1 Percentage of valid points inside buffers with different distances

We can observe that the AHC variation had a larger difference between the two levels of dissimilarity than the KDE+AHC variation in which the two levels were almost similar. This could be explained due to the use of percent volume contour in the KDE+AHC variation. During this process we used the 95% of the points that were used to generate the kernel density estimate, so small clusters with few observations and individual vectors were not included. By this way, we were avoiding the weakness of AHC algorithm which is the sensitivity to outliers. This weakness is also more intense at the lowest levels of dissimilarity. But when we calculated the clusters in the next level of dissimilarity, the AHC variation had higher percentages than the KDE+AHC variation. That means that the kernel density, combined with the percentage volume contour was a restrain for the KDE+AHC variation, especially when we reached at higher values of percentages of validation points in the different buffers i.e. at more optimal number of clusters. These percentages were higher for AHC variation in all the different radii of buffers e.g. in radii 10 meters AHC variation had 76.64% validation points included but KDE+AHC had 72%, in 20 meters 87.53% for AHC and 82% for KDE+AHC and so forth.

As we mentioned before the optimal cluster configuration can be recognized only by subjective interpretation and highly depend on the application (Jung, Park, & Du, 2003). In the KDE+AHC variation, the number of spatial clusters calculated from the kernel density process was 39. In order to calculate also the temporal characteristics of these clusters and to define the optimal spatio-temporal clusters, we had for each one spatial cluster to perform the AHC algorithm and define separately the number of final spatio-temporal clusters. In the AHC variation, we configured only once the number of spatio-temporal clusters and we also avoid further configuration of the Kernel density estimator at specific cases that it might be necessary.

In Table 3.3 we summarize the above results and present the advantages and the disadvantages of each variation.
Table 3.3 Comparison of the two methodology variation

<table>
<thead>
<tr>
<th>Methodology Variation</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>KDE+AHC Variation</td>
<td>1. Solves the weakness with outliers</td>
<td>1. Time consuming</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2. Lower percentage of validation points</td>
</tr>
<tr>
<td>AHC Variation</td>
<td>1. Faster implementation</td>
<td>1. Sensitivity to outliers</td>
</tr>
<tr>
<td></td>
<td>2. Higher percentage of validation points</td>
<td></td>
</tr>
</tbody>
</table>

For the evaluation that is following and for further analysis of our data, in which we include the environmental features and the activities of pedestrians, we used the AHC variation. The use of AHC algorithm only, was proven more flexible and simpler than the first one, and requires comparatively less time to be executed. The tests showed that the results of both variations (“KDE+AHC” and “AHC”) are equivalent. However, the second method has more advantages, it is simpler and does not need a parameterisation of the search distance in which is the case of the “KDE+AHC” variation. Therefore, we used the results from the AHC variation for further analysis.

3.3.4 Defining the number of clusters

The definition of the final number of cluster after the implementation of AHC algorithm is very difficult. We used the evaluation data in order to define the optimal number of clusters for the experiment 2010. In Table 3.2 we observe some characteristics of the two different dissimilarity levels of AHC variation. The average duration per stop is 1.8 minutes at AHC test 543 and it is increased by 136.1% at AHC test 159. The number of stops per person decreased from 29.7 to 15.75 (15.53 at the control dataset) and the average number of persons per spatio-temporal cluster was also increased from 1.53 to 2.77.

In Figure 3.2 we can see the differences between the two tests in the value of persons inside the spatio-temporal clusters. In the case of the AHC test 543 we can observe that 67.4% out of 543 spatio-temporal clusters had only one person inside them. On the contrary in AHC test 159 this percentage was only 17% out of the 159 spatio-temporal clusters. The clusters which include 2 persons inside were 26.7% in test 543 and 55.3% in AHC test 159. The percentages with more than two persons inside were lower than 10% with the only exceptional the case with cluster with four persons inside where the test 159 has 13.2%.

The design of the experiment was that the participants should walk in couples. In Figure 3.2 we can see that in AHC test 543, more clusters include only one person but in AHC test 159 more clusters include 2 persons, as how the experiment was performed. Also there is a slightly increase of clusters which include 4 persons in AHC test 159, but not in AHC test 543. This corresponds to the control dataset, where more than one couple stopped at some places.
Figure 3.2 Spatio-temporal clusters with specific number of persons within them

Figure 3.3 depicts a good example of the importance of the aggregation level. In the first aggregation level, several small clusters are produced, whose mean centres are scattered around the locations of the control dataset (i.e. the mean centres of the pictures). In the second aggregation level, those clusters are merged, and the locations of their mean centres better correspond to the control dataset.

Figure 3.3 Test 543 and Test 159. Example of mean centres of individual stops and mean centre of the set of pictures

We also tested the results using different levels of dissimilarity and assessing the number of control stops inside increasing distances around the mean centres (Figure 3.4). In the test 543, for example, 50.36% of the control stops were inside a buffer of 10 meters, and this percentage increases slower at larger radii. Similar tendencies were found in the other tests. However, in test 159, the angle from 1 to 10 meters is steeper reaching a value of 76.64%. At 20 meters the percentage is 87.53% and at 40 meters is 90.7%. After the comparisons between the different tests, which represent different dissimilarity levels, we conclude that the less spatio-temporal clusters that we had, the less the percentage of validation points we had inside the different buffers, a proportional relationship was formulated. This was happening until test 159 and then an inverse proportional relation was formulated, where the larger the number of spatio-temporal clusters we were calculated, the less the percentage of
valid points we had inside the buffers of the mean centres. The choice of test 159 is the optimal configuration of clusters.

![Graph showing percentage of valid points inside buffers with different distances](image1)

**Figure 3.4 Percentage of valid points inside buffers with different distances (AHC Variation)**

The steeper angle that is presented in test 159 is reasonable after a certain distance, since there will be more individual stops with more different distances from the validation points. This result also provides the optimal distance to compute the buffers for further analysis, i.e. 20 meters. At this distance, 87.53% of the validation stops were detected. In Figure 3.5 the above decision is confirmed. The increase of the percent of validation points (% of total) that occurred at the distance of 20 meters was 4.31%. This increase was larger in all the different distances before the 20 meters (with only exception the 15 meters) and in all the distances after the 20 meters the increase was less than 4.31%.

![Graph showing increase of percentage of valid points inside buffers with different distances](image2)

**Figure 3.5 Increase of percentage of valid points inside buffers with different distances**
3.4 Temporal Patterns

3.4.1 Introduction

For the definition of temporal patterns we used the dataset with the GPS recordings from the Experiment 2006. Implementing the AHC variation, we calculated 1,850 spatio-temporal clusters and 1,997 individual stops in the lowest level of dissimilarity. 123 of the spatio-temporal clusters (6.65%) had between one and four visitors. The rest of the spatio-clusters included only one visitor. 733 (36.7%) individual stops included a single vector. We assigned to those stops a duration of 15.14 seconds (the average time between two consecutive GPS records). 194 (9.7%) of the individual stops had more than 10 movement vectors with an average duration of 5.53 minutes. In Figure 3.6 the frequency distribution of the number of stops per visitor is depicted. The average number of stops per visitor was 5.82 and the average duration of each individual stop was 1.18. The average stopping time was 6.88 minutes (6.04% of total time they spent in the park).

![Figure 3.6: Frequency distribution of the number of stops per visitor](image)

We run a test with a higher level of dissimilarity at the AHC algorithm i.e. with less number of spatio-temporal clusters (test 1543) as we did in the evaluation process. An example of these results is depicted in Figure 3.7, where dots represent movement vectors for one visitor and the large circles represent vectors classified as suspension. Each individual stop is represented with different colour. The same movement suspension vectors were clustered as one individual stop at the test 1543 and as three different cluster at the test 1850 which is the one that we are using for our results. The longest distance between these vectors was approximately 158 meters. Also the first GPS vector was recorded at 08:36:23 AM and the last vector was recorded at 08:40:55 AM, a difference of 4.53 minutes. In between the vectors that were characterized as movement suspension from LISA process, there were intervened movement vectors. With this test, we could confirm that the lowest level of dissimilarity (test 1850) in this case is more optimal than the next level of dissimilarity (test 1543).
3.4.2 Distribution of stops in different periods

We analyse the results by comparing the occurrence in different time periods, such as weekdays and weekends and for May and August. We can observe in Figure 3.8, that there is an increase in the total number of stops during the weekends of August. At May the total number of stops was 945 and in August 1052, representing a small increase of 5%. The difference between weekdays and weekends in total was 13%. The largest difference was between the weekdays and weekends of August (increase of 26%).

If we compare the results of the number of stops in Figure 3.8 with the total number of visitors in these periods (Figure 3.9) then we could observe that they are proportional. The more persons visited the park at that period the more stop were formulated. In the number of visitors there was also an increase at weekends, especially in August, 104 visitors.
The average durations of the individual stops remained relatively stable. Neither the different months nor the different parts of the week seem to influence these values. The lowest value was in weekdays of August, 0.95 minutes and the highest value at Weekdays of May, 1.29 minutes.

The average numbers of stops per visitors also showed not significant variation (Figure 3.10). The visitors stopped in average 10.28% more times in August than in May and 6.3% more during the weekends than in weekdays. The largest value is at weekends in August comparing with all the other categories.

The difference at the average numbers of stops between weekends of August and weekdays in August is 17.36% and the difference between the weekends of August and weekends in May was 19.67%. The largest total number of stops per month is in August, 6.22.

In Figure 3.11 is depicted the average stopping time as a percentage of the total time that visitors stayed in the park. The largest percentage of stopping time is at May and at weekdays, 7.48%. Comparing the two months, August has less stopping time than May. The lowest stopping time was at the weekends in August 4.75%.
The number of stops that occurred at the weekdays of May is lower than the average number of stops at all the periods (Figure 3.8). We already mentioned that the average duration of the stops remained relatively stable during that date periods. The large percentages of stopping time at the weekdays of May could be explained from the total time that the visitors stayed in the park, which was the lowest of all the other periods, 107.32 minutes. On the contrary the low values of stopping time in weekdays in August could be related with the number of stops in this time period which is the lowest of all, in combination with the values of average duration per individual stop, which is also the lowest.

In Figure 3.12 the number of encounters is depicted as a percentage of the total number of stops. We could observe that the largest percent of encounters was at weekends of May, 11.99%. The lowest values were at weekdays of August when it reached 1.71%.

The low proportion of encounters at the weekdays of August could be combined with the number of stops at that period which was the lowest and with the average duration of stopping time in the park which was also the lowest at that period. So we can associate the number of encounter with the number of stops and with the duration of the total stopping time inside the park. Moreover, strong associations between the different date periods and the stops that we calculated are not formulated.
The other periodicity that we analyse is in different hours of the day, regardless the date period, month or weekdays. If a stop occurred between two different time periods then it was calculated as a stop for both of them.

In Figure 3.13, we can observe that the largest number of stops of all the visitors during the park was 440 stops during 11:00 and 12:00 and the smallest at periods 1 and 8. We performed the same categorization but with the number of visitors per hour and we found that the results are proportional with the number of stops (Figure 3.14). We can see that the largest amount of visitors (165) is at time period 4. The smallest numbers of visitors were at time periods 1 and 8.

In Figure 3.15 we compare the average duration (in minutes) of stops with encounters and stops without encounters, in each time period. For the individual stops without encounters, the largest duration was during time period one, 1.65 minutes. All the average durations of the stops without encounters were calculated within a range of 0.78 until 1.65 minutes. We can observe that the duration of stops without encounters didn’t have large differences between the time periods.
On the contrary the average duration of the stops with an encounter had major differences between the different time periods and it was larger in all of the time periods than the stops without an encounter. The average duration of all the time periods for the stops with encounter was 4.01 minutes. The largest duration was in period 7 when the stops with an encounter had an average of 7.18 minutes. These stops, 16 in total, occurred at specific locations in the park. 7 of them (43.75%) occurred at the Sheep farm near the visitor centre and 3 of them at the information panel of this sheep farm. Sheep farm has the largest average duration of individual stops of all the park features as we analyse in the next chapter.

In Figure 3.16 the number of encounters as a percentage of the total number of stops per time period is depicted. The largest number of encounters per stops occurred at period 6, 9.39%.
To have a better perspective of the encounters that occurred at the study area, we calculated also how many stops with an encounter (percentage of total stops) occurred at the different sequences of the stops. In Figure 3.17 we could observe that there is an increase of encounters at the largest number of stop sequence.

![Figure 3.17 Individual stops with encounter (% total number of stops) that occurred at each specific sequence of stops](image)

The more encounters occurred at the stops towards the end of the visit in the park. This is also obvious after combining these results with the results in Figure 3.16 in which the largest percentage of stops with an encounter was at time period 6. These periods were not the periods with the more visitors inside the park. Instead there were the periods that the stops with an encounter have the largest durations (Figure 3.15).

### 3.5 Geographical Elements (Destinations, connectivity of paths)

Using the optimal distance identified in the validation experiment of 2010, we computed buffers of 20 meters around the individual stops that we calculated from the experiment 2006 and identified 25 geographical features in the park related to them. In the cases that there were not any park features in the distance of 20 meters around the individual stop, we also included the intersections of the paths. We also added at the categorization of the features, 8 more categories to have a better perspective of pedestrians’ stops at these features. 2 of these categories were “Unknown” and “Undefined” as they are described in Chapter 2.5. The other 6 categories are related with some specific amenities inside the park (Figure 3.18). The individual stops that occurred at these locations, have included more than one park feature. For example, in the entrance of Parking 1 there were also 3 picnic tables and one information panel.
Specific amenities where stops occurred with more than two park features included

The stops that occurred at Parking1 and were including one or more of the park features that we just mentioned were categorized as “Parking1+“. These categories were formulated for 4 parking lots, 1 Snack bar positioned at northeast of the park and the Radio-telescope. At the Radio-Telescope, all the individual stops that we calculated were associated with more than one and at maximum 13 park features, which was also the maximum number of features associated at all the individual stops.

The individual stops that were not related with any park feature or path intersection were 386 (19.32% of total) and they were categorized as “Unknown”. This category was the most frequent category in the dataset. They were covering the second largest number of vectors of the whole dataset, since these individual stops are including 1152 movement suspension vectors (12.81%) but an average of 2.98 vectors per individual stop. That means that these stops that occurred at the “Unknown” category were stops with the smallest duration of all the categories, 0.8 minutes, very few vectors per stop less than the average of all the stops and also occurred across the whole study area. The stops at this category also have the largest number of distinct visitors. 180 visitors (52.47%) stopped during their residence in the park at these areas.

The visitors entered and leave the park from 5 different parking lots. Four out of five parking lots have more than one park feature related at the same area. These park features were picnic tables, benches and information panels.
In total, 494 stops (24.73%) were located at the parking lots. 342 of these stops could be also associated with some other feature such as picnic benches or information boards. The average duration of these stops was 1.08 minutes, significantly higher than the average duration of 0.67 corresponding to all the stops related to parking lots. 59 encounters were detected at the parking lots, 41 of them corresponding to parking lot 1. The Visitor Centre Natuurmonumenten next to the Parking lot 1 was associated to 141 individual stops, 7.06% (Figure 3.19). The average duration of the stops at this place was the 2.59 minutes, the second largest after Sheep farm with average duration 5.46 minutes (Figure 3.20) and the number of encounters was 26 (18.84%), also the second largest.

Figure 3.19 Number of individual stops per park feature

Figure 3.20 Average duration of individual stops per park feature
We also found three more interesting places associated to large number of stops: the Radio-Telescope, the Sheep farm which is located approximately 600 meters from the Visitor Centre Natuurmonumenten (Figure 3.21) and the benches across the park. The first two features, Radio-telescope and Sheep farm, have 69 and 38 individual stops respectively. These numbers were not as large as the number of individual stops from the features we mentioned before, but they have relatively high average of duration at these individual stops. For the first feature, it was 2.01 minutes and for the Sheep farm it was the largest of all the park features, 5.46 minutes.

Figure 3.21 Radio-telescope, Visitor Centre and Sheep Farm

Furthermore, all the individual stops that occurred at the Radio–telescope were including at least two more park feature. In a distance of 65 meters from the Radio-telescope, there were 5 picnic tables, 7 benches, 1 ANWB Mushroom and 1 information panel. At the Sheep farm, 12 encounters (8.69%) occurred at the stops.

In the whole park there were 92 benches. 127 individual stops occurred at these benches with an average duration of 0.96 minutes per stop. These stops were covering 394 movement suspension vectors (4.3%) and the average number of vectors per individual stops was 3.1. The number of the unique visitors that visited at least once a bench during their residency at the park was 89 (25.9%) the fifth larger from all the park features.

One more kind of feature that we took into consideration is the path intersection. There are 1864 different paths around the park with a total length of 296.75 km. The total number of points in which these paths were intersected is 1236 locations. We assigned at the individual stops a path intersection only if there was not any other park feature related at this position. The individual stops which occurred at a path intersection were 208 (10.41%) with an average duration of 0.63 minutes. The number of vectors that were included at these stops was 556 (6.18%) and the number of vectors per individual stop is 2.67.

Summarizing the results we could recognize some specific patterns that were formulated from the pedestrians’ movement inside the park. They are still a lot of individual stops that
are categorized as “Unknown” i.e. without having any park feature associated with them. We assume that more spatial characteristics should be taken into consideration, such as the topography of the area, the visibility from these positions, to have a more coherent perspective for these stops. The parking lots were the feature from where the visitors entered and left the park, so it is reasonable to have higher values at the above characteristics than the other park features. The number of stops to the main amenities of the park such as the Snack bar, the tea house, the lookouts and the Sheep Farm were less than the average number of stops which was 62. The Radio Telescope had slightly higher, 69 individual stops. The average duration of the individual stops for all the amenities in the park was 1.10 minutes. An interpretation for these characteristics could be that most of the visitors have first priority to walk or to hike in the park and not to stop at specific amenities.

### 3.6 Activities

The result of the questionnaire provides a list of the activities of pedestrians during their visit in the park. From the 461 participants at the survey only 343 (74.4%) actually carried a GPS. The question that referred at the activities was “What is the main activity of today (1 answer)” and there was a choice between 11 different answers. There was also another category in which there were including two or more different activities. All these categories are depicted at Figure 3.22. 247 visitors (72.01%) answered that they visited the park for “Walking/Hiking”, 26 visitors (7.58%) had more than one activity. At this category, 23 visitors included walking as the one of the activities and the other activities were one or more of the predefined answers of the questionnaire or something else e.g. observing flora/fauna, searching mushrooms. 19 visitors (5.53%) answered dog walking and 19 answered other than the above answers. All the rest of the categories were covering 28 visitors (8.19%) and 4 visitors didn’t answer the question.

![Figure 3.22 Number of visitors per activity](image)

In Figure 3.23 we can see, how the visitors used their time during their residency at the park. We found that the longest stopping time corresponded to visitors whom the main activity was “Taking Pictures”. We assumed that the movement suspension vectors were related with individual stops inside the park and all the rest of the vectors are related with moving from one location to another. The activity with the largest percentage of time that the
visitors spent in stops was “Visiting Sheep Farm” and it was 22.79% out of the total time. The next larger percentage of stopping time was in the category of the visitors that had not answered at the specific question (21.94%).

The visitors that were taking pictures during their residency in the park were the third largest percent, spending 17.71% of their total time for stopping around the park. The next activities with the higher percentage of time spending in stops were “Going to the visitor centre” with 15.75%, “Picnicking” with 11.75% and “Observe Flora/Fauna” with 10.52%. The category “Walking/Hiking” was having a relatively small percent of time spending in stops which was 4.92%. The visitors that answered this activity were covering 5,600 movement suspension vectors (62.31%) and also during their activity had a sum of 94 encounters in total 68.12% of all the encounters occurred in the park.

The visitors that were taking pictures have the largest average number of individual stops which is 11.4 stops per visitor (Figure 3.24). The activity with the second larger average number of individual stops was “Observe Flora/Fauna” which had 8.78 stops per visitor.

Combining the results of Figure 3.23 and Figure 3.24 we could come to the conclusion that the activity “Taking Pictures” were having a largest duration of stopping time due to the average number of individual stops per person but also due to the average duration of this individual stops (2.26 minutes). Specifically, the visitors were stopping more times for taking pictures than in the other activities. We could observe the same tendency for the activity “Observe Flora/Fauna”. The third largest average number of stops was 6.33 for the activity “Visiting Sheep Farm”.

Figure 3.23 Average duration of non-stopping time and stopping time
We also combined the park features with the activities of the pedestrians. The visitors that answered that their main activity was “Walking/Hiking” had in total 1467 individual stops which were covering all the park features that were mentioned in chapter 3.5. From these stops, 298 occurred at a position marked as “Unknown” (20.18%) and 166 individual stops (11.25%) occurred at path intersections. For the visitors that were at the park for “dog walking” the number of individual stops was 53 and the two more frequent park features that these stops took place were “Unknown” and Parking 4. Visitors that answered the category “Other” for main activity, in which there were included answers such as family visiting, walking, sightseeing had 12 stops at the Visitor Centre (12.5%) and also 12 stops to “Unknown” attraction. Visitors that were taking pictures during their residency in the park had 57 individual stops in total and from them 23 of them were at “Unknown” attractions and second more frequent park feature was the wetlands with 6 individual stops (10.52%). Also for the visitors that were observing flora and fauna at the park they had 79 individual stops in total. Although the three most frequent park features (45.56% of all the stops) that these stops occurred were at the visitor centre, path intersections and the category “Parking lot 1+”, the benches were covering 11.4% of the stops and wetlands the 8.8%.

The activities of the pedestrians were not directly associated with their stops but with their goal of the visit. Nevertheless, we can recognize that during their stay in the park their activities were related more with recreation than with commercial activities and the park features had been directly associated with these activities.

4 Conclusion and Recommendations

4.1 Conclusions

This study builds upon the use of GPS/GIS technology in analysing pedestrians’ movement suspension. Although Movement Suspension Patterns (MSPs) have been studied before, the problem lies in the lack of temporal analysis of these phenomena. Our contribution is to
expand the analysis of MSPs not only from spatial perspective but also from temporal in order to improve the analysis of movement suspension. Our results answered the questions that we developed for the purpose of this research.

For the temporal analysis, we aggregated the computed stops in different time periods. There were not significant differences between the monthly periods, May and August. Furthermore, we found that the number of visitors does not determine the amount of stops with encounter, but only the number of stops. In fact, encounters are usually associated to stops with larger durations (the mean duration of all the individual stops was 1.18 minutes but for the stops that encounters occurred was 4.45 minutes).

In order to include the geographical elements of our study area with the individual stops, we focused on the cross-paths and on park features. In general, we found that 81% of the individual stops were associated with one of those features. The rest of the stops were not associated with any of the park features. We can conclude that the commercial destinations of the park, Tea House and Snackbar Spier were not visited often by the pedestrians (lower number of individual stops from the total average number of stops). The visitors showed a preference for some park features such as the Radio-Telescope (high number of individual stops) and the Sheep Farm (high average duration of individual stops). However, most stops in our dataset had relatively small duration (71.5% of stops had duration less than one minute) and occurred in locations that none park feature could be associated to them or locations across the park with benches. Combining this with the results of the activities, the visitors preferred places that could take pictures or could have some rest after walking or hiking.

We included in our research the survey that conducted in the experiment of 2006. We attempted to relate the activities reported by the visitors with the detected stops and the associated park features at these stops. We were able to identify some general characteristics of the stopping behaviour (number of stops, duration) with the activities reported by the visitors in the survey. For example, we found that visitors reporting that their main activity in the park is to take pictures or observe flora and fauna had a large number of short-duration stops. On another example, visitors whose main activity was to visit a specific attraction of the park, the number of stops is small but with longer duration.

In order to retrieve the above answers to our questions, we used the approach of Orellana and Wachowicz (in press) to extract movement suspension patterns from GPS tracking datasets and we applied two different clustering techniques to analyse the temporal dimension of the detected patterns, KDE+AHC and AHC variation. After the evaluation that we performed, we conclude that the AHC variation has significant advantages over the KDE+AHC variation i.e. faster implementation and higher percentage of validation stops (87.53% of validation stops were included in the distance of 20 meters for the AHC variation instead of 82% for the KDE+AHC variation). Another advantage of using the AHC variation is that the only necessary parameter is the number of clusters based on the outcome of the algorithm. The evaluation that we performed confirm that our methodology is adequate to analyse and explain the formulation of the spatio-temporal patterns. It could also facilitate
to expand the analysis of human movement by including geographical elements and the activities of pedestrians.

4.2 Recommendations

One of the most crucial parts of our methodology is the number of clusters that we defined as an output after the implementation of AHC algorithm. Once a cluster is merged the next step will be based on the previous cluster to calculate the newly generated clusters. Thus merge decisions may lead to variable results (Han & Kamber, 2006). The problem of the optimal decision of the number of clusters is not new. There have been developed already various methods for defining the optimal configuration or the evaluation of the decision (Jung, Park, & Du, 2003) (Milligan & Cooper, 1985). One suggestion can be to combine different clustering techniques to improve the clustering quality and then evaluate the results. Clustering techniques that could be combined with AHC algorithm is BIRCH or ROCK (Han & Kamber, 2006). With the controlled experiment, we were able to define the optimal number of clusters for the analysis of the stopping behaviour of visitors of Dwingelderveld National Park using GPS tracking data.

We also consider that it is necessary to include for further analysis other environmental components (e.g., safety, aesthetics) and take into consideration the assumption of researchers that the visibility factor of environmental features influences the pedestrian behaviour. They support that the route and destination choices are partially influenced by the landscape visually perceived along the street network, beyond purely functional aspects of movements (Foltête & Piombini, 2007). For analysing human activities under a spatial and temporal framework in other context than visiting natural areas, it is also important to collect additional information such as socioeconomic and demographic characteristics, constraints of scheduling activities. There are already researches which support the assumption that the context of these activities is important. The timing of the activity, the kind of place where it is pursued, the persons involved and various factors associated with the personal and geographical background of the individual initiating the activity also affect the degree of spatial and temporal fixity (Schwanen, Kwan, & Ren, 2008).

The detection of movement patterns can be used to optimize the design of location-based services. The services offered to a moving user could not only be dependent on the actual position, but also on the estimated current activity, which may be derived from a detected movement pattern (Gudmundsson, Laube, & Wolle, 2008). For analysing human movement on that level, ethical and privacy issues should also be taken into consideration (Dobson & Fischer, 2003). Furthermore, our methodology can be applied as a tool in different scientific fields such as traffic management, animal behaviour and urban planning.

5 References


Bentley, R., Jolley, D., & Kavanagh, A. M. (2010). Local environments as determinants of walking in Melbourne, Australia. Social Science & Medicine, 70, 1806-1815.


