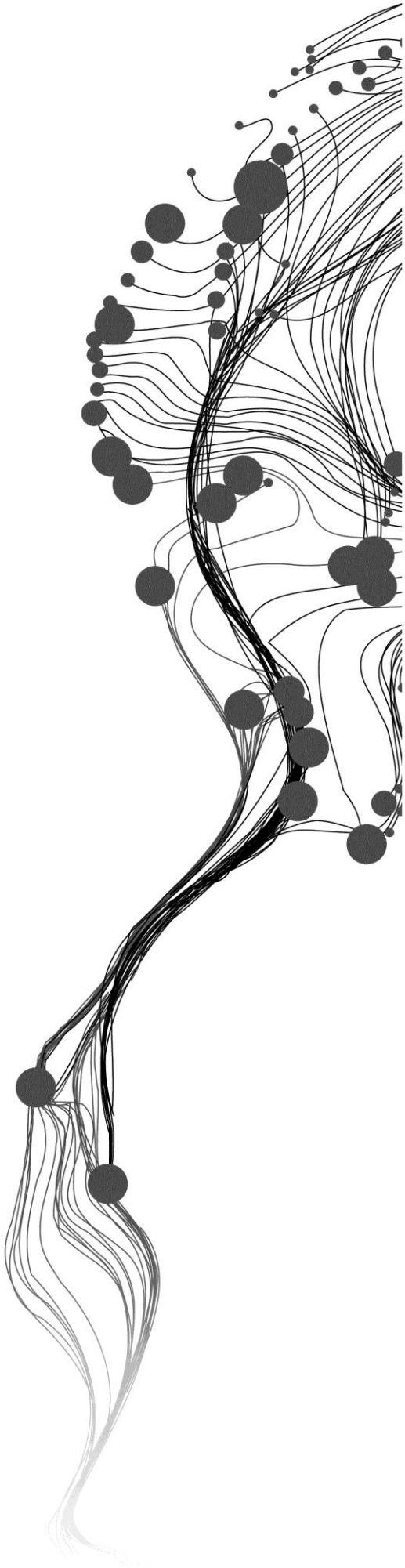
A CLUSTERING-BASED APPROACH FOR ENRICHING TRAJECTORIES WITH SEMANTIC INFORMATION USING VGI SOURCES

PARYA PASHA ZADEH MONAJJEMI
February, 2013

SUPERVISORS:
Dr. U.D. Turdukulov
Drs. B.J. Köbben



A CLUSTERING-BASED APPROACH FOR ENRICHING TRAJECTORIES WITH SEMANTIC INFORMATION USING VGI SOURCES

PARYA PASHA ZADEH MONAJJEMI
Enschede, The Netherlands, February, 2013

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.
Specialization: [Name course (e.g. Applied Earth Sciences)]

SUPERVISORS:
Dr. U.D. Turdukulov
Drs. B.J. Köbben

THESIS ASSESSMENT BOARD:
Dr. A.A. Voinov
Dr. O. Huisman, ROSEN Inspection & Services, Germany

ABSTRACT

The vast collection of movement data through various mobile devices generates a significant and precious amount of information that can provide valuable insight into movement behaviours in various application domains.
Different studies have been done towards interpreting this data using standard spatial queries; in order to have a better understanding of their nature and dynamics. Yet there still is a limited but growing amount of work that considers the semantic of the movement that links the spatio-temporal data to its context. Hence there is a necessity to address this concept in relation to mobility data.
Analysing the movement data in association to its background geography requires a workflow that can determine the important parts of a movement track and link them to the significant places in the background. This will be more effective if performed considering the application domain and the data characteristics. In this research different methods were studied and examined in order to develop this type of workflow and to annotate the important parts of movement tracks with their corresponding background information. The effect of different variables (mainly speed and density) was studied on a pedestrian movement dataset in order to extract stops from the individual trajectories. This was followed by stacking up these results using a raster overlay to determine the significant places which are indicating higher importance in the background geography. The discovered places were attributed by relevant descriptive information extracted from online crowdsourcing web applications (mainly Foursquare). And finally the spatio-temporal tracks were annotated, and additional information was attached to the points using spatial join between the track points and extracted polygons of places where the moving object has stopped or slowed down.
The resulted polygons of significant places were not always in the same size and same level of accuracy (larger polygons in the crowded centres resulted from slow movements of many people compared to smaller more detailed ones on sparse regions). Also Foursquare as a data repository could not provide precise information for all the queried positions.

Keywords: Trajectory data, density based analysis, trajectory semantic annotation, spatio-temporal studies, VGI, crowdsourcing

ACKNOWLEDGEMENTS

I am extremely grateful for all the amazing time that I could have spent and all the valuable knowledge and experience that I have gained during my stay in The Netherlands.

First of all I wish to offer my utmost gratitude to Dr. Ulanbek Turdukulov, my first supervisor who has always supported and guided me throughout this research with his patience and knowledge, providing me freedom in thinking and developing ideas, while leading me toward the best directions. And Drs. Barend Köbben, my second supervisor whom I could always have the most fruitful discussions and receive constructive comments about my work during the research. I consider myself so lucky for having the chance to work with such great people. I also wish to extent my thanks to all GFM professors whom I have learned a lot from each one of them during my studies.
I must also acknowledge Dr. Otto Huisman, for offering this research area in the first place, and for helping me to develop an understanding of the subject and choosing this topic.

It is a pleasure to thank all the nice people that I have met here, for making every moment joyful and invaluable for me. I would love to thank my best friends or my *Indian sisters* Zoya Sodhi and Aroshaliny Godfrey for their presence in my life and for their absolute, boundless friendship. For all the moments we have shared and for all the memories we have made together. I also want to give thanks to my great friends Manuel G. Garcia and Ahmed Eissa not only for their precious friendship and support, but also for everything that I have learned from them so far. I feel so blessed that too many names are coming to my mind to be thankful for knowing them. My sweet GFM classmates, my friends back home, my friends from all over the world! I wish I could thank them all one by one…

Finally, I want to express my deepest gratitude to my dear family. My mother Minoo, my dearest one, my closest friend, my sweetest advisor, the one that without her I could never had reached where I am now. And my brother Shahriar for cheering me up whenever I need and for being the one that I can always count on.
It will be my honour to dedicate this thesis to my dear father. The one, who has always believed in me, and has truly supported me during my decisions and actions in the most significant and crucial moments of my life. My debt to him is uncountable.
.

TABLE OF CONTENTS

LIST OF FIGURES

1. INTRODUCTION

1.1. Motivation and problem statement

Today, mobile devices are widely available, and the enormous growth of their usage produces a massive amount of mobility data. Presence of mobile phones, wireless sensor networks, devices equipped with Global Positioning Systems (GPS) and Radio-Frequency Identification (RFID) causes an explosive growth of geo-located data which enables tracking the movements by a low cost continuous capture of points in a time sequence ("GeoPKDD," 2013). This data flood suggests an emergence of algorithms and analysis to explore, understand, model and exploit this type of data.

The results of these studies will provide an insight to the human behaviour by tracing their movement, discover their manner and habits and their centre of interest. Information and knowledge about how and why people are moving is needed for scientists, planners and managers for taking informed and intelligent actions in different fields from urban planning and traffic engineering to commercial purposes and social sciences (Giannotti & Pedreschi, 2008).

Trajectory data, representing movement data or mobility data, is usually generated as sequences of *id*; *x*; *y*; *t* points through mobile devices (Bogorny et al., 2011). This data is required to be processed into more human-perceptible structures in order to facilitate further analysis. In a conceptual model suggested by Spaccapietra et al. (2008) these are referred as *episodes* of stops and moves. Stops are considered to be locations where the moving object shows a different behaviour like slowing down or stopping at the same place for a period of time, episodes are usually defined based on speed threshold or staying in a certain distance from a particular location. Whereas G. Andrienko et al. (2011b) suggests an event based view of the movement and aims for extraction of *movement events* or *m-events* with relevance to the goal of analysis and tasks.

Deriving patterns using standard spatial queries or spatiotemporal operations on raw trajectory dataset is not always sufficient for an interpretation of movement behaviour and building knowledge. However, generalizing the observations and linking them to their context by adding semantic information, will help to gain insight into the dynamic processes. Enriching trajectories with semantic information is effective to simplify the queries, analysis and mining the movement data, in different application domains (Alvares et al., 2007). Current projects as GeoPKDD[1] and MODAP[2] emphasize on the necessity of addressing the semantic concept related to mobility data.

In the present thesis, concept of semantic trajectories refers to the sequences of episodes with geo-located information about POIs (Points Of Interest). POIs are application dependent and they can places, stations, mountains etc. Here the main interest is to find places of common interest according to pedestrian movement behaviour. For instance, in a semantic trajectory a stop will be known as a Café, instead of Coordinates of 52.22, 6.89 (G. Andrienko et al., 2011b; Spaccapietra et al., 2008).

The data type, representing movement, is usually massive and complicated. Aggregating or classifying the trajectories, and analysing group behaviours instead of individual tracks are among possible approaches towards recognition and interpretation of patterns in movement data. One example of this type of

[1] Geographic Privacy-aware Knowledge Discovery and Delivery – http://www.geopkdd.eu/

[2] Mobility, Data Mining, and Privacy ("MODAP," 2013) – http://modap.org/

approach can be the discovery of common points of interest in the whole dataset based on combination of individual POIs, extracted from single trajectories.

The semantic information about the places, that connects movement data to its context, can be derived from different sources. Among the possible data sources are cadastral maps, questionnaires, and crowdsourcing geospatial data or VGI (Goodchild, 2007). VGI sources especially can be a beneficial option to consider for this purpose (Heipke, 2010).

As a key feature of Web 2.0 this type of data is uploaded by users on the internet through different services. Various social networks and portals like Wikipedia, Flickr, Twitter, Blogs, and mobile applications like Instagram and Foursquare are motivating people to provide significant amount of geolocated data in different forms.

Thus far, existing approaches have not focused on developing an efficient workflow to automate this type of processes in an application domain. Linking the trajectories to their environment by defining and annotating episodes and extracting significant places (common POIs) based on discovering patterns in groups of trajectories is suggested to expedite intelligent analysis over individuals or groups of moving objects.

In similar experiments (Van Langelaar & S.C., 2010), this has been basically done by overlaying the data on base layer maps and matching interesting points with their semantic information manually to understand trajectory patterns in relation to their background.

Current research is oriented to problems in movement data analysis that require that require delineating places of interest based on movement characteristics in relation to its spatial context (G. Andrienko et al., 2011b)

These procedures and methods are not always efficient, generic and scalable; an automated approach is required to ease this process and to have a higher efficiency and speed. Hence general design architecture to facilitate this process from user's point of view is requisite.

1.2. Research Identification

1.2.1. Research objectives

Main objective

- To design a workflow for automatically enriching the trajectories with semantic information.

Sub objectives

- To discover episodes and POIs in trajectory dataset, using clustering methods.
- To extract location information about discovered places from the user generated content, accessible through internet (i.e. crowdsourcing geospatial data or VGI).
- To integrate extracted background information with the trajectories in order to produce semantically annotated trajectories

1.2.2. Research questions

For the workflow design:

1. Which components are considered in designing an application workflow and how do different specifications affect the effectiveness of the workflow?
2. How generic can this workflow be and for which types of movement data can it be used?

For discovering places:

1. Which methods can be used to extract interesting parts of trajectories (i.e. stops)?
2. Which methods can be used for grouping the individual stops and discovering common POIs?

For extracting location related semantic information

1. What are the available VGI sources and which one fits the research's requirements best?
2. How to extract related information from the selected VGI sources?
3. Can a combination of VGI sources be used to cover the possible gaps in one of these sources?

For annotating the trajectories and analysis of movement patterns:

- Which methods should be used for integrating the semantic information with raw trajectories?
- What geometry type on trajectories and what level of granularity should be considered for the annotation?

The innovation of this research is aimed at developing a new workflow (framework) for automating the process of discovering significant places from trajectories' own characteristics, finding attributes off those places and annotating trajectories with this semantic information in order to enable further analysis of group activities.

The synthesis for the discovery of POIs and also the information sources (VGI) used for annotation step are the main focus of innovation for this work.

1.3. Thesis structure

The remainder of this thesis is structured as follows. Chapter 2 starts with some basic concepts and definitions and continues with providing the literature review on the most important aspects of the research along with basic information about related methods and techniques applied in related work. Chapter 3, describes the suggested procedure, methodology and experiments. The results for each main step of the research are also presented at the end of corresponding methodology sub-section in the same chapter. Chapter 4 starts with results of map overlay to validate the obtained results, this is followed by discussions about the methodology and results and is completed with recommendations about presented work. The last chapter (Chapter 5) includes conclusion and future work.

.

2. LITERATURE REVIEW

Before starting the literature review section, it is helpful to present some definitions and phrases, and to describe basic concepts in the following sub-section, as these will be repeated frequently in this dissertation

2.1. Basic concepts and definitions

Trajectory data

Movement data (also referred as mobility data) (G. Andrienko et al., 2011b) collected by mobile devices, can be represented either by continuous (Hägerstrand, 1970), or discretized list of $(id; (x_N, y_N, t_N))$ in which t represents the time intervals while (x_N, y_N) are associated spatial coordinates of the points acquired at t_N. In a discrete representation each travel has an id and N number of track points (a finite set of observations) that form the trajectory and a motion in space and time (Spaccapietra et al., 2008).

This type of data representation only has location information and geometrical patterns. In order to be able to add more human-comprehensive attributes to the trajectory and link the movement to its concept it is required to produce conceptual trajectories.

In this thesis regardless to the definition of *Trajectory* by Spaccapietra et al. (2008) which refers to segmented GPS tracks as trajectories, the words trajectory and track are used to address the data collected as $(id; (x_N, y_N, t_N))$.

Episodes of Stops and Moves

The conceptual model for trajectories suggested by Spaccapietra et al. (2008) allows the user to add semantic information to the particular segments of movement data. In this model trajectories are converted to structured recordings of movement organized in Episodes of Stops and Moves, in between start and end (Figure 2-1). The type of information added to the trajectory is dependent on the type of movement and the nature of moving object. For example for a bird's fly path, the added semantic information can be the length of stay in a particular place or the temperature at a certain time, while for a human movement it this additional attribute contains information like the places in which he stops or the roads that he passes.

.

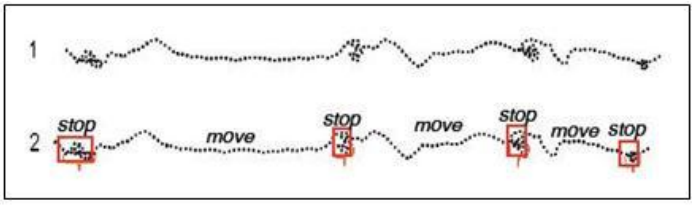


Figure 2-1: Stops and Moves in a trajectory (Spaccapietra et al., 2008)

In the literature review section, along with various existing approaches towards adding semantic information to the trajectories, some of the suggested methods for defining and extracting episodes and stops in different application domains are also mentioned.

Common POIs:
Nonetheless, not all personal stops are considered as significant places to the application domain, for example in an urban study about pedestrian movements, the location of one's house and office has a lower degree of importance compared to public places and common POIs; even though these types of locations will be extracted as stops in single trajectories. Therefore, discovering repeated stops or frequent visits, or common behaviour among various trajectories that are related to the same geographic entity is significant (Figure 2-2).

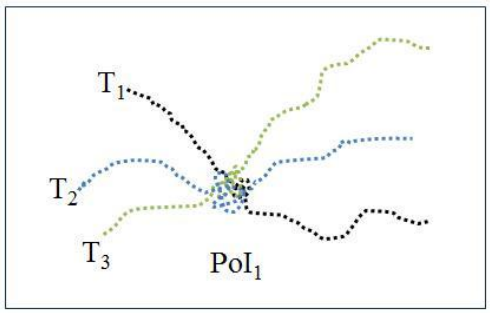


Figure 2-2: Common POI

Considering the nature of studied data and the application domain targeted in this approach, personal stops will be neglected and will be annotated as *unknown stops*. While discovering places in the context that are more important to majority of moving objects will be defined and the corresponding individual stops will be annotated by the attributes of these places (i.e. pedestrians in the city centre visiting touristic spots).

2.2. Related work

A general overview of existing approaches and techniques for processing and analysing movement data is provided by G. Andrienko et al. (2011a). However there are still limited studies that take the semantics of trajectories into account.
Clustering methods have been widely used in data analysis and data mining applications, for discovering highly correlated regions of objects by abstracting the underlying structure. These methods have been widely applied on spatio-temporal data or moving object trajectories, to discover patterns and interesting behaviours for extracting information.
Applying clustering methods over all the trajectories (instead of single tracks) in order to discover certain patterns of movement (e.g. meet, flocks, convergence etc. (Dodge et al., 2008)) is also an alternative approach that can lead to uncover the location of interesting places. Density based clustering algorithms are used to find moving clusters of objects and to extract movement patterns (e.g. sequential pattern, frequent pattern, association rules) in (Kalnis et al., 2005; Laube et al., 2005).
Whereas in (G. Andrienko et al., 2011b) through visual analytics, after extracting the episodes, density based clustering was used to delineate relevant places considering their position in space and time and also additional attributes such as thematic values. And it properly deals with cyclic attributes.
Among various studies towards understanding, processing, mining and analysing movement data, those which follow the idea of structuring the trajectory into segments of stops and moves (episodes), have more relevance to the purpose of this research. As this model enables dealing with the movement data as individual trajectories and the results for each individual can be used for discovering group behaviours. Therefore, mainly the follow ups for Spaccepietra's model are presented in the current section.
Having presented a conceptual model for trajectories, Spaccapietra et al. (2008) did not focus on the ways of extracting stops and moves. This approach was further taken by Alvares et al. (2007) presented the SMoT (Stops and Moves of Trajectories) algorithm. In this algorithm, an application dependent list of places (regions) is required as an input for defining candidate stops. The trajectory might have a stop in

any of these candidate stops. SMoT matches every single point in trajectory with every candidate stop, in order to extract stops and moves. The polygons in which the moving object spends specific *minimum time duration* are labeled as stops; whereas Moves are defined segments between two stops or one stop and the beginning or end point of a trajectory.

The biggest disadvantage of SMoT is the requirement of pre-defined POIs for extracting stops. This issue was further taken to consideration in (A. T. Palma et al., 2008) presenting an alternative to SMoT named CB-SMoT (Clustering-Based SMoT). This algorithm operates in two steps; first it uses a variation of DBSCAN algorithm (Ester et al., 1996) (a density based clustering) to identify potential episodes, considering additional parameter of *speed*. Second, it matches these episodes with the geography behind the trajectories considering the duration of stay. In this approach list of pre-defined places is not required for the geometrical extraction of stops. While for the annotation with semantic information and linking the data to its geographical background the POIs are taken from the user.

Current research advocates a similar approach to achieve the first objective (discovering common POIs),. Clustering based approach was selected and applied on data as track points to extract the stops and was followed by discovering the location of common POIs.

A brief description of different clustering methods and relative literature using these methods is provided in this section to justify the selection of density based calculations in this study. According to purpose of analysis and targeted objectives.

Clustering algorithms are mainly defined in two main groups: *Partitional* and *Hierarchical*. Data mining clustering techniques are also grouped into: Density-Based Clustering, Grid-Based Clustering and Model-Based Clustering. (Kaufman & Rousseeuw, 1990)

Partition-based algorithms, divide the database D of N objects into a set of K clusters of objects. K is an input parameter chosen by user, depending on his knowledge about the data. This algorithm starts with an initial partition from the data and applies an iterative control strategy to optimize the results of objective function for each cluster.

Each cluster is specified by its gravity centre (k-means) or by one of the objects that is closer to the centre (k-medoid) (Jain & Dubes, 1988). CLARANS (Clustering LARgeApplicatioNS) (Ng & Han, 1994) is the first clustering algorithm designed for Knowledge Discovery in Databases (KDD). In this algorithm, given an input parameter k, k random objects from the dataset are chosen as cluster seeds and all other objects are assigned to their nearest seed. The results are then refined by moving objects from one cluster to the other as long as achieving a stable configuration. Therefore, every point is assigned to one cluster. High complexity and not being able to produce arbitrary shaped clusters are two main disadvantages of this method. This method was adopted for studying human mobility in (Giannotti et al., 2011) by applying an overall clustering in the trajectory data mining.

Hierarchical algorithms create a hierarchical decomposition of objects in a dataset *D*, represented by a dandogram. The dandogram divides *D* to smaller subsets with an iterative approach. This approach can be agglomerative (bottom-up) or divisive (top-down).

An *agglomerative* approach or a bottom-up merging, start with taking each object as a cluster itself, and further merge groups according to a distance measure. This grouping will stop after each object is placed in a cluster, or when the user wants. As opposed to the agglomerative algorithms, the *divisive* approach starts with one group of all objects and successively divides this group, into smaller ones. But similarly It will stop either by the will of user or when each object is placed in a cluster.

In one of the earliest approaches towards this aim, Ketterlin (1997) models trajectories as sequences of points and considers a conceptual hierarchy over the elements of this sequence. Whereas (Nanni, 2002) applied 'k-means and 'hierarchical agglomerative clustering methods' on trajectories.

Grid-Based clustering algorithms mainly focus on spatial data, and their objective is to quantize the dataset into cells. Even though these methods are similar to hierarchical approaches, merging of clusters is decided by a predefined parameter and it is not based on a distance measure.

STING(STatisticalINformation Grid) (Wang et al., 1997), WaveCluster(Sheikholeslami et al., 1998) and CLIQUE (CLustering in QUEst)(Agrawal et al., 1998) are three representatives of this group of algorithms.

Density based clustering algorithms, agglomerate objects within clusters, based on density. Or in other words, the basis is the population within a given neighborhood in space. It groups objects according to a defined *density objective function*. Two parameters that define a density based clustering are Minimum Number of Points (n_{pts}) and a given radius of Epsilon (ε) in a way that for each object in the same cluster, the neighbourhood of the radius ε has to contain at least n_{pts}objects. Each cluster will grow until the minimum number of points does not exceed the n_{pts}. Unlike partitional algorithms, Density-based algorithms do not use iterative allocation of points to a given number of clusters. Hence, these algorithms are robust to noise and outliers

Among the density based clustering algorithms, DBSCAN and its follow ups are the most highlighted methods that are briefly going to be described.

DBSCAN (Density-based Spatial Clustering of Applications with Noise) (Ester et al., 1996) is the most well-known density based clustering algorithm and it has influenced many other density based methods up to now. This algorithm defines clusters as the set of points that are transitively connected by their neighbourhood.

Similarity is the main implication in clustering algorithm, when the criterion for this similarity is selected as *distance* between the dataset elements; it results in nearest neighbour clustering (NNC). DBSCAN looks for the core clusters and expands them by aggregating nearest neighbours meeting certain conditions.

DBSCAN needs two parameters; *Eps*and *MinPts*. *Eps*is the parameter for delineating a neighborhood whereas *MinPts* is a density measure that indicates the minimum required amount of points in a neighborhood with *Eps* radius in order to assign that point and its neighbors to a cluster. And if C_1, C_2, …, C_N are clusters in a dataset D with respect to *Eps* and *MinPts*, the set of the points in D that do not belong to any $C_i$are *Noise*.

DBSCAN is an advantageous algorithm from different aspects, it can be applied to complex data at reasonable computational costs; it finds arbitrary shaped clusters and is not sensitive to the input order. Besides, every newly added point can only affect a certain neighborhood; therefore the overall quality of results will not be reduced. This algorithm was enhanced later in following research works in different ways according to application and requirements and to enhance efficiency and reduce the amount of calculations. The main deficiency of this method is that it requires minimum number of points and the neighborhood distance as input parameters from the user, which itself requires a pre-knowledge about data.

Density-based clustering was oriented towards geographic data in (Sander et al., 1998) by introducing GDBSCAN that is developed based on DBSCAN.

Another variation of DBSCAN is introduced by Lee et al. (2007) performed on line segment clustering for clustering the trajectories.

DJ-Cluster (Li et al., 2010)(Density and Joined based Cluster) is another algorithm created based on DBSCAN that uses a concept of connected components instead of the connectivity notion of a clique graph, therefore it enhances DBSCAN's performance.

DJ-Cluster has been used for discovering personal gazetteers and it has been more precise than k-means method. This method attaches semantic information to the clusters but it also requires user to provide list of important places, then it checks if these locations are placed in the trajectory.

ST-DBSCAN (Bamis & Savvides, 2010)(Spatial-Temporal DBSCAN), is another algorithm developed based on DBSCAN, but it is able to treat non-spatial attributes as well as spatial ones. This is done by using two parameters for distance measures, ***Eps1*** and ***Eps2***, The latter is used as a similarity measure for not spatial attributes.

On the other hand, Kisilevich et al. (2010b) presents P-DBSCAN, a density based clustering algorithm, applied on geo-tagged photos. This method introduces "density threshold" and "adaptive density for fast convergence towards high density regions" as novel concepts (Kisilevich et al., 2010b, p. 2).
As it was mentioned before, CB-SMoT is formed based on Spaccepietra's model for trajectories and it uses an enhanced clustering method developed based on DBSCAN. In addition to CB-SMoT, different variation of SMoT algorithm are also presented for extracting similar information, taking various parameters into consideration. Among these algorithms IB-SMoT an Intersection-Based and DB-SMoT which is Direction-Based spatio-temporal clustering (Rocha et al., 2010) can be mentioned.
Zimmermann et al. (2009) applies an enhanced algorithm developed based on CB-SMoT on error-prone recording devices and considers a threshold for duration of stay as well and is more suitable for the data that the speed is not homogenous (e.g. the GPS tracks while using different transportation modes)
Another extension of CB-SMoT is presented in (Idrissov & Nascimento, 2010) that partially automates the requirement of initial parameters, in this implementation from *MinTime* and *Eps,* the second one is not required to be defined by user, instead an area unit is used by the algorithm.
OPTICS (Ordering Points To Identify the Clustering Structure) (Ankerst et al., 1999), is an evolution of the basic DBSCAN that is much less sensitive to input parameters and offering results as a reachability plot with an efficient selection of input parameters. This algorithm computes an augmented *cluster ordering* for interactive and automatic clustering of the data. The clustering structure is represented by this ordering, and it contains information that is equivalent to a clustering that is obtained by a range of parameter settings.
Among these methods density based clustering fits the best to the requirements and purpose of the current research, the reasoning is further explained in more details in section 3.3.1.
A different approach towards extracting and interpreting the significant places in movement data is taken in (Gennady Andrienko et al., 2007), in which visual analytics tools are used for this purpose. Gennady Andrienko et al. (2007) extract the position of stops from the dataset based on a threshold for the time interval that is spent in each position. Furthermore the extracted places are visualized on to overlay them on a map display as point symbols. For extracting the repeated stops (i.e. common POIs) OPTICS algorithm is used for the spatial clustering of casual stops on close positions or overlays
Cao et al. (2010) propose a framework for extracting meaningful locations from GPS tracks, in which locations are also ranked according to their significance based on duration of stay, frequency of visit and the distance user travels to reach there. The algorithm used in this work is SEM-CLS which is an enhanced algorithm compared to OPTICS (Ankerst et al., 1999) and K-means (Jain &Dubes, 1988). Clustering algorithms and the main indicators of them that had an influence on the current research are briefly presented in (Figure 2-3).

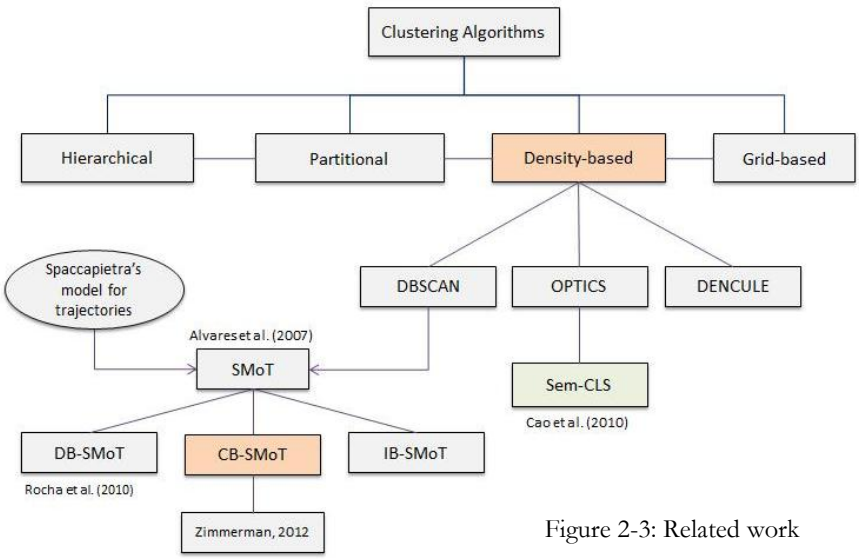


Figure 2-3: Related work

On the subject of clustering the movement data, one recent work (Orellana et al., 2012) suggests a new approach. Orellana et al. (2012) introduces a method that applies a Local Indicator of Spatial Association (LISA) to find spatial clusters of low speed vectors of movement data, in order to "detect movement suspensions of patterns for collective objects". This method is applicable for different movement data; it is scale-independent and does not require spatial and temporal thresholds.

Moreover while similar methods of clustering and analysis are applied to user generated contents in (Hong et al., 2012; Kisilevich et al., 2010a; Kisilevich et al., 2010b; Zhou & Meng, 2011) this type of content (i.e. VGI) can be used as an alternative source of background information for annotation of trajectories.

At the current moment the user contribution to web and mobile applications is extremely high and the amount of geolocated data produced through volunteered geotagged updates is massive and is continuously increasing (Naaman, 2011). Considering the fact that commercial services may have a lot of financial and legal limitations. The idea of using an alternative for these services by using the user-generated output of different applications as an input for others is promising. For the aim of this research the accessibility and availability of the data in addition to the type of required information will be considered to select the most suitable platform or a combination of them.

Annotation of trajectories with semantic information is done in (Spaccapietra & Parent, 2011) and (Alvares et al., 2007) following the discovery of interesting places. Yan et al. (2011) presents a framework (SeMiTri) for annotating heterogeneous trajectories, here the episodes are annotate in three layers: region, line and point of interest (ROI, LOI, POI), it is a generic framework applicable to various types of trajectory data.

The three basis of this design are “latent motion context”, “layered approach” and “heterogeneity of semantic places”.

Relevant annotations are selected based on the context of the movement and depending on the application type. Annotations are attached to the episodes of trajectories instead of track points. This is to avoid information overload by aggregating the correlated records

This approach uses a spatial join for annotating the trajectories with information about the regions (ROIs). While map matching is used for the road networks (LOI), and hidden Markov model for inferring semantic points (POIs).

Placing a selection of all these methods in a proper workflow still promises possible enhancements since very limited work is available on this subject in the literature (Bogorny et al., 2011; Yan et al., 2011).

3. METHODOLOGY AND RESULTS

3.1. Data description

The main dataset used for this research, is a dataset containing traces of time stamped GPS locations of pedestrians, moving in the central part of city Delft, The Netherlands. This data was collected in November 2009 by TU Delft /Urbanism for a European project called Spatial Metro ("Spatial Metro," 2012). The data is also available within the MOVE project (COST IC 0903) upon request ("Move-Cost," 2012).

The "Tracking Delft 1" dataset includes ±300 GPS tracks with the frequency of 2 seconds for track points, collected in 4 days in .gpx format. Among these, 292 tracks were accepted as valid in the dataset provided for this research. This validation was based on readability and consistency of tracks, existence of the date of collection and also a valid match with the questionnaire for each track.. GPS devices were distributed among the participants in access points to the city and every participant made one trip. In addition to the tracking data, mandatory questionnaires were filled by the participants in the data collection process for a qualitative feedback. These are stored in a database (MDB file) and they include various fields of information about owner of the track (i.e. age, gender, occupation, purpose of trip, familiarity (frequency of visits) and weather conditions).

This data was converted to .csv format using GPSVisualizer[3] to ease further processes and analysis, since this format is compatible with most of the available software for spatial analysis.

3.2. Overview

This work can be considered as a design-research thesis, in which an inductive approach is going to be followed for designing a new synthesis of existing methods with modifications with respect to the application and data type, in order to achieve a more efficient automated flow of techniques (Trochim, 2006).

After extensive literature review on the related topics an initial model of workflow includes three major steps:

1. Geometric pattern discovery
   - Extracting stops on individual trajectories and
   - Discovering common POIs over the whole dataset.
2. Extracting related background information from online sources and
3. Semantic annotation of trajectories.

Each of these steps is going to be described in details in the following sub-sections of current chapter. At the end, an enhanced and completed workflow design will be formed based on the initial sketch. This improved workflow will be realized in a prototype that receives raw GPS tracks as input data and produces semantically annotated trajectories with minimum contribution from the user. Figure 3-2 illustrates the initial flow diagram for these steps in a general overview.

[3] http://www.gpsvisualizer.com

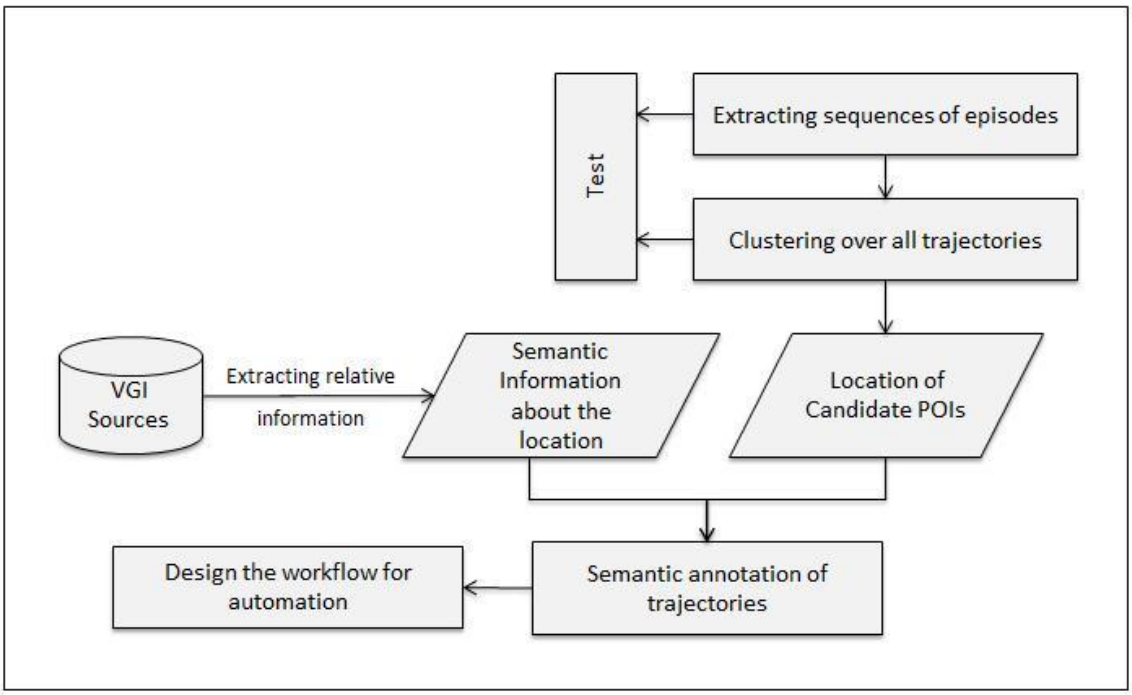


Figure 3-2: Methodology - Workflow

3.3. Geometric pattern discovery: Extracting common POIs

The first phase of the current research suggests a methodology for discovering the significant places or common POIs in a moving objects dataset, from the characteristics of raw trajectories. This is achieved in based on extraction of individual stops and analysis of common behaviours among the group of trajectories.

Discovering common POIs from the geometrical characteristics of the dataset without considering the background geography, is done in two steps that are presented in the following subsections.

3.3.1. Extracting stops on individual trajectories

With respect to the literature review (Spaccapietra et al., 2008), (Alvares et al., 2007), (A. T. Palma et al., 2008), as a first step, trajectories were processed to produce structured trajectories that have episodes of stops and moves. This was done by extracting the potential stops or important parts of the trajectory as sub-sequence of points in the original track.

Among the existing methods that were discussed in the literature review for structuring a trajectory into stops and moves, a clustering based approach was decided to be followed in the current methodology as this research was aimed to extract stops without any dependency on the geographic background knowledge.

This indicates that a pre-defined list of potential stops is not available and methods like (Alvares et al., 2007) could not be applied for extracting potential stops. But with a similar methodology to (A. T. Palma et al., 2008), a clustering approach was selected to be followed for discovering potentially important parts in individual tracks.

The considerations and justifications for selecting *most appropriate clustering method*, *additional criteria* and *proper values for input parameters*, are described in the current section.

As described in section 2.2, deciding on one particular clustering algorithm as the most effective and efficient method is not possible, since depending on different criteria as data characteristic (dimension and size) or the objective function, one specific method can work well on one dataset while being very poor for another one. But some of the characteristics of a good clustering are: scalability, finding arbitrary-

shaped clusters, minimum requirements as input parameters, handling of noise, sensitivity to the order of input records, interpretability and usability.

Having known these characteristic and comparing different methods with respect to our dataset, for the following reasons, density based clustering algorithms were suggested for the implementations in this research work:

- Number of clusters is not needed to be known in advance.
- These algorithms treat sparse regions as noise and are robust with respect to noise. Since trajectory data often contains underlying random components and/or low resolution of the measurements, it is highly important to choose an algorithm with noise tolerance.
- Results will be generated in arbitrary shapes of clusters (like hierarchical method but with a considerably lower complexity ($O\ (nlogn)$ vs. $O\ (n2)$).

Despite the general strengths of density based algorithms compared to the other methods, application dependent reasons and justifications were also taken into consideration for selecting this method over others. Some of these justifications are listed as follows:

- In a moving object dataset, higher density of points in an area unit (e.g. a region), represents higher duration of presence in that region, as the GPS device keeps recording points as far as the object is staying at that same location and the number of track points increases by time.
- As noted before, it is not necessary for density based clustering algorithms to have the number of clusters as an input parameter. Therefore an unknown input parameter will be omitted from the workflow requirements.
- Higher frequency of visit in a region is another reason for having more track points in one location. Hence if the moving objects doesn't stop or stay at a location but visits a certain location more than a time, it can still be concluded that this region possible has an importance for the further analysis (e.g. main squares and junctions for urban studies and city/traffic planning)

On the other hand, only a *higher density* is not a sufficient indicator and measure for defining stops in a trajectory when actual physical stops are targeted. In other words a location with a higher density along a trajectory might be significant but since it does not represent a stop in the movement, extraction of that location will be a false detection.

To avoid getting this type of results another criterion had to be added to the applied method. The challenge was to select proper trajectory attributes to use for this purpose. According to the literature, some of these attributes are duration of stay in a place (Alvares et al., 2007), travelled distance, direction of move (Rocha et al., 2010) average or maximum speed etc. . Similar to CB-SMoT algorithm, *speed* was the chosen parameter for discovering the stops in individual trajectories.

In the proposed methodology, before applying a clustering method on an individual trajectory, the points of each trajectory were filtered based a speed threshold. This threshold was applied on the instant speed attribute of points. Hence those points that were considered too fast for representing a stop were omitted from the trajectory before applying the density based clustering.

After filtering the positions based on speed, "point density calculation" was applied over the remaining points to identify the densest *region*s over the area that the object was traveling. The point density method requires *population field* and a *neighbourhood definition* (including shape, size and area units) as inputs. When there is no population field selected, this method counts the number of points in an area of $A = \pi r^2$ (where r is the selected radius as an input parameter for defining the neighbourhood), and calculates the $D = N_{pts}/A$ where D is the density and N_{pts}, represents number of points. The output for point density calculation is the number of points per square unit area. This area unit was selected as meters for easier and more meaningful analysis and comparisons with real world.

The proper values for these input parameters were initially set after exploring the dataset and using common sense. Later after numerous empirical comparisons between different values on various tracks, final parameters were fixed for the final calculation model that was applied on the entire dataset.

For the speed parameter, the initial value was set to 0.7 m/s considering the nature of movement data (i.e. pedestrian movement) and logical reasoning. According to studies(Knoblauch et al., 1996), average speed of pedestrian walking, varies between 1.1 m/s and 1.6 m/s depending on various factors (e.g. age, gender, activity type, etc.). Therefore, by applying this threshold, only points were selected where the pedestrians slowed down. In addition to the initial value, more diversity of values was examined to confirm the proper speed threshold. After comparison between (0.7, 0.6, 0.5, 0.4 m/s) speed threshold was set to 0.6 m/s. This selection has filtered the points in which the object is clearly not stopped. But it still preserved those that might belong to slowdowns (Figure 3-3).

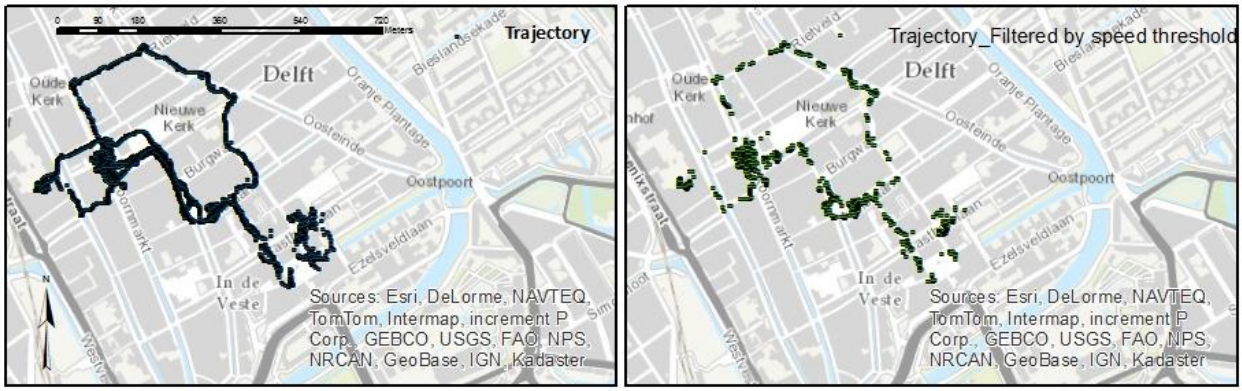


Figure 3-3: Filtering points based on speed threshold of 0.6 m/s

However this speed limit was selected rather high, therefore false points could have still stayed in the selection but it would have provided a confidence that all the possible stops are discovered. In addition, other type of significant places that the pedestrian might slowdown and be wandering around them, but not actually stop there, are also going to be discovered with this limit.

The neighbourhood for the point density calculation was defined as a circle with the radius of $r = 5\ m$. This was decided after comparison between 5, 10, 15 m on different trajectories. A comparison of neighbourhood definition with $r = 5\ m$ and $r = 10\ m$ for a single track is presented in Figure 3-4. The smaller the radius the more precise and dense are the results for particular places of interest. In the contrary when the neighbourhood is defined with extreme small radius (i.e. 2-3 m) large areas get omitted. Whereas, $r = 10\ m$ generates larger areas with lower density. The result of this step was the raster with different density values over each trajectory boundary.

This calculation was then followed by a reclassification of produced raster into two groups in order to mark the results as stops where it meets a selection criterion. Stops were defined as the classes that their density value is higher than d. This value was selected as $d = 0.4\ points/m^2$ based on logical reasoning and experimental comparisons. As the GPS device has the temporal accuracy of 2 seconds, ideally in each minute it should record 30 points (If no point is cleaned or lost during that time).

The nature of places which are considered as potential POIs varies from small local shops or cafes to big shopping malls, an average area of a potential POI was supposed to be 75 m² which also stands as the

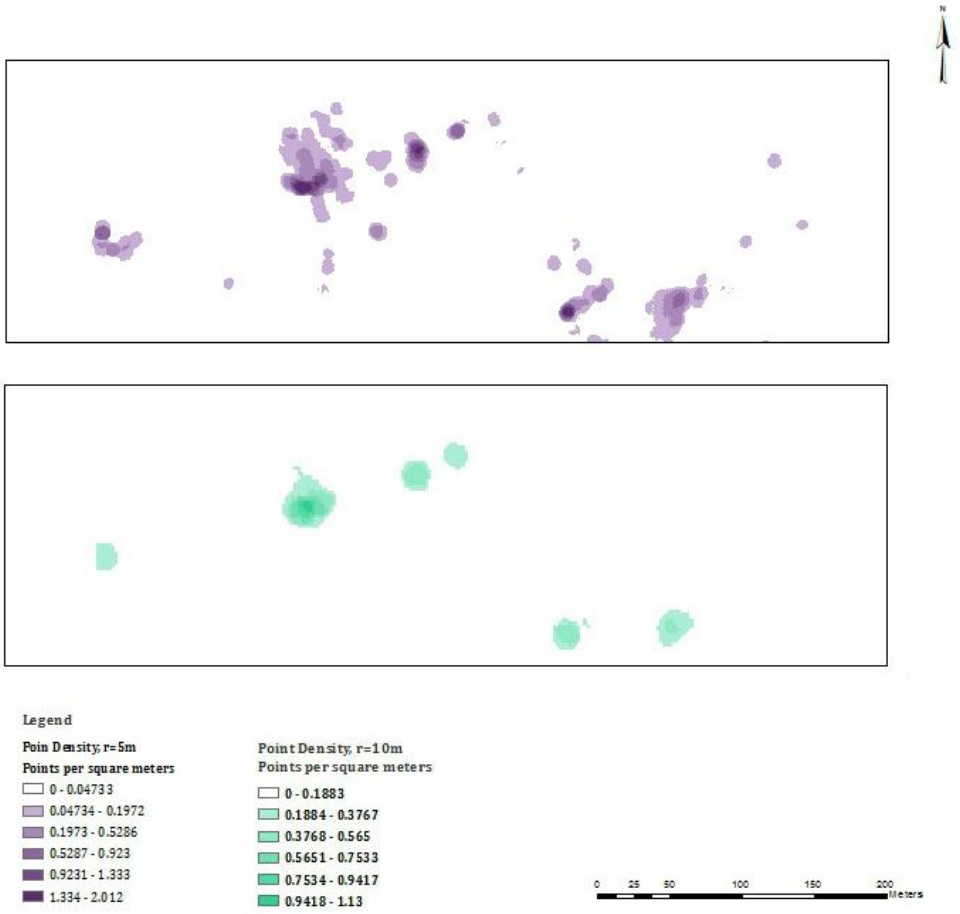


Figure 3-4: Point density calculation 5 vs. 10m

same area as the defined neighbourhood for density calculation. Therefore if the moving object stays or stops in a place with area of 75m² for a minute the point density of that place will be 30/75 m² or 0.4 point/m². At this point *the duration of stay* from (Alvares et al., 2007) and (A. T. Palma et al., 2008) corresponds to the selection of 1 minute, and 30 points that are supposed to be acquired in that period of time.

Thus the value of 0.4 *points/m²* was selected as the threshold for defining stops in the raster result of point density calculation. This reclassification was applied on results of point density calculation for r =5 m and r = 10 m to analyse the results before finalizing the selected values. As r =10 m produces rather large areas (the area of the calculation neighbourhood is 314 m²) with relatively lower densities (compared to r=5m) many results are getting omitted and the remaining one are big areas with the same density spread around an area of 314 m². Therefore, neighbourhood with r =5 m was chosen since this threshold omitted the regions with lower density and left more meaningful and precise areas.

In the final result of the reclassification the two new classes are standing for stops, with a density higher than 0.4 *points/m²*, and the rest of points (start, moves and end) are stored as NoData and are omitted from the raster. The output cell size was set as 1m considering the necessity of level of details and the size of output raster images (Figure 3-5).

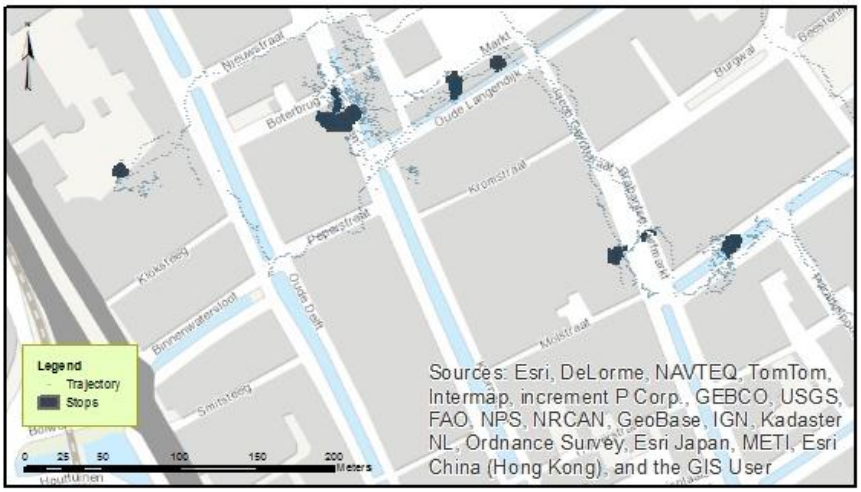


Figure 3-5: Reclassified individual stops

This process was repeated for all the trajectories following the method presented in Figure 3-6. An important remark is to define the same extent for all the trajectories through all the calculations.
These classes were then converted into polygons for further use in the upcoming steps of the methodology.
In order to have the results of this implementation in both raster and point format, each track was overlaid on its corresponding polygon of stops and the points of stops were selected based on their topological intersection with these polygons. For each tracks all the points that are falling into a polygon of stop are labelled as stops if their speed is less than 0.6 m/s^2. The rest of the points will be representing moves.
Eventually these parameters are corresponding with the two main parameters of DBSCAN algorithm, *MinPts* and *Eps*. *Eps* parameter is equivalent for the radius of the neighbourhood area as the diameter of this area would be the maximum distance between two points belonging to one circle. And these circles are going to be connected if they have an equal or similar density values. Whereas *MinPts* can have the same meaning as *d,* the selected value for the reclassification.

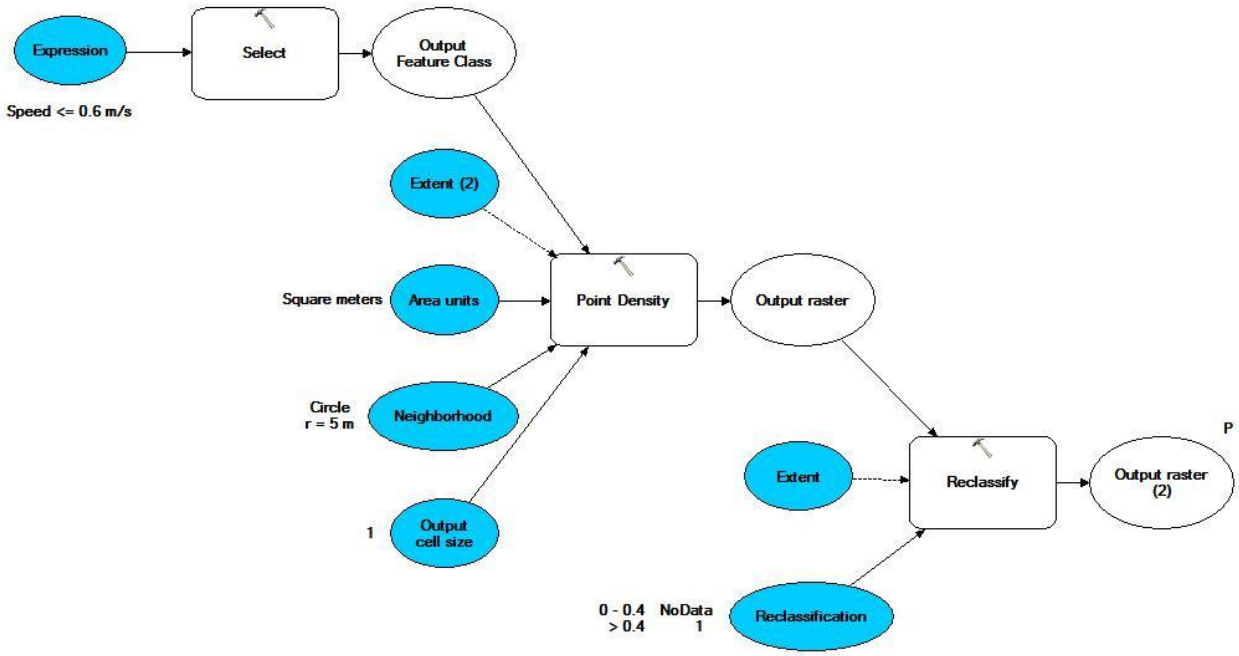


Figure 3-6: Process - Extracting individual stops

3.3.2. Results

The results of this step are:

- The raster data for each trajectory, representing stops.
- The matching polygons with these rasters.
- The corresponding points to the stops on each trajectory.

The presented model was applied over 292 GPS tracks and the results were stored as rasters and points. Among these tracks 42 of them did not have any enough dense regions to be an indicator of stop so they were automatically omitted from the further analysis (during the reclassification based on the threshold of 0.4).

3.3.3. Discovering common POIs over the dataset

Having extracted the stops for individual trajectories, in the second phase the process was followed by discovering common stops or common POIs. These individually extracted stops can be meaningful locations with importance for the application purpose or they can be personal preferences of the moving object like his home. It is important to define the places that more than one track is passing from there and more than one person shows interest in those locations.

At this stage it seemed to be possible to treat the results of first step as either raster data or points, as we prepared the data in both formats. These two alternatives were taken into consideration and the raster was selected for the follow ups, but the points approach is also going to be discussed at the end of this section. In the suggested method a raster overlay of stops was made to extract the areas that are corresponding to stops and are common among more than a certain number of trajectories. The output of this process was a new raster with 21 different values; meaning there were some areas in the region that more than 20 trajectories had stops in that region. Each value represents the number of previously extracted raster stops. (Figure 3-7)

Figure 3-7: All stops, raster overlay

This raster was reclassified in order to have a homogeneous and meaningful result that represents common stops. Here, all the values higher than 3 were placed into one new class with value 1 were as the rest were labelled as NoData.

At this stage the produced raster was cleaned and noises were removed. Morphological operations and raster cleanings (post classification filters) were applied and examined in order to find a best way which omits small areas and cleans the boundaries of bigger shapes. At the end a 3*3 filter of expand followed by shrink were used to clean the ragged edges between zones (Figure 3-9 and Figure 3-8)

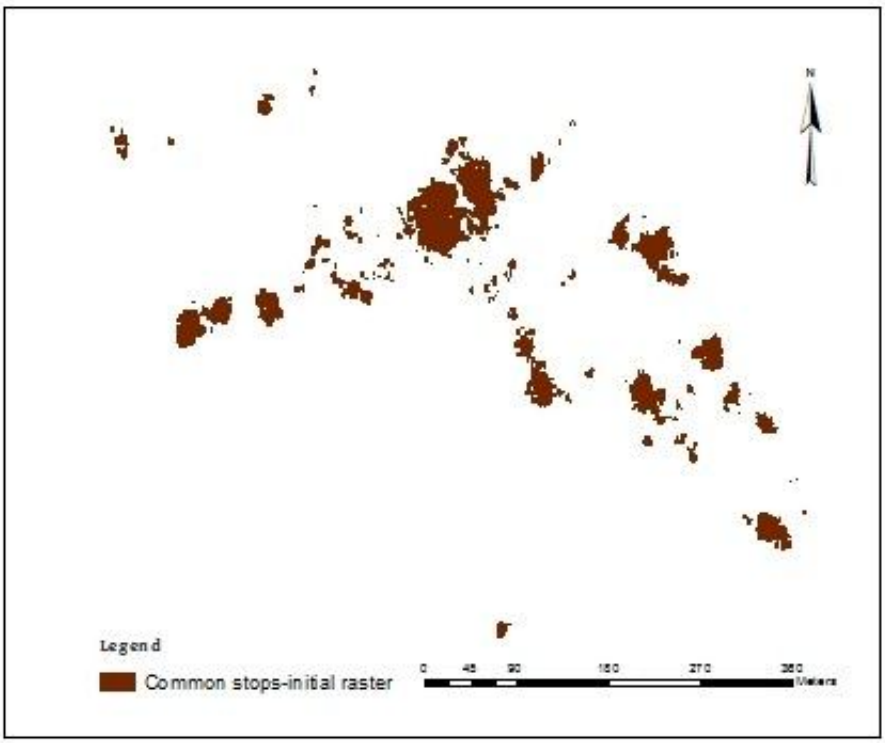


Figure 3-9: Common stops – initial, reclassified raster

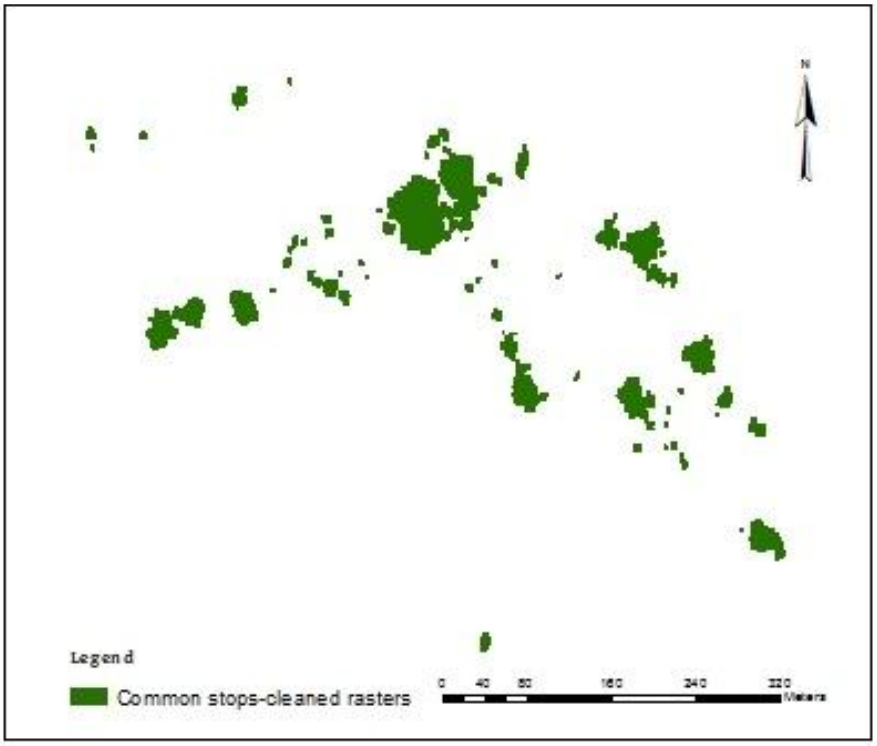


Figure 3-8: Common stops - cleaned, reclassified raster

After this step the refined raster was converted to polygons and the coordinates of the centre of each polygon was extracted and saved as an additional attribute to these places as *Points* of interest (Figure 3-10, Figure 3-11). These locations were stored in order to be used as an input parameter to extract related attributes around this location from online data sources.

OBJECTID	grid_code	....	area	center	center_y_	Lat,Long
....	....	....	....	....	....	....
45	1	....	1125.757	4.360637	52.01028	52.0102824716581,4.36063694322479
46	1	....	21.72591	4.362668	52.01003	52.0100311823971,4.36266751892972
47	1	....	15.43549	4.363348	52.01	52.0100035067872,4.3633481308303
48	1	....	12.34841	4.362634	52.00991	52.0099106449275,4.36263425528327
49	1	....	919.3747	4.362223	52.0101	52.0101031143132,4.36222291784097
50	1	....	210.5006	4.363914	52.00988	52.0098823168717,4.36391415021063
51	1	....	30.87116	4.362745	52.00973	52.0097307250742,4.36274452874179

Figure 3-10: Table of attributes - polygons, centers

As a second alternative an overlay of all the points was considered as a possible option to find the dense areas and extract them as polygons again. This approach produced a new raster calculated by point density method applied over all the points that belonged to stops. This process did not produce the same results or regions with the targeted meaning (i.e. common between more than two tracks).

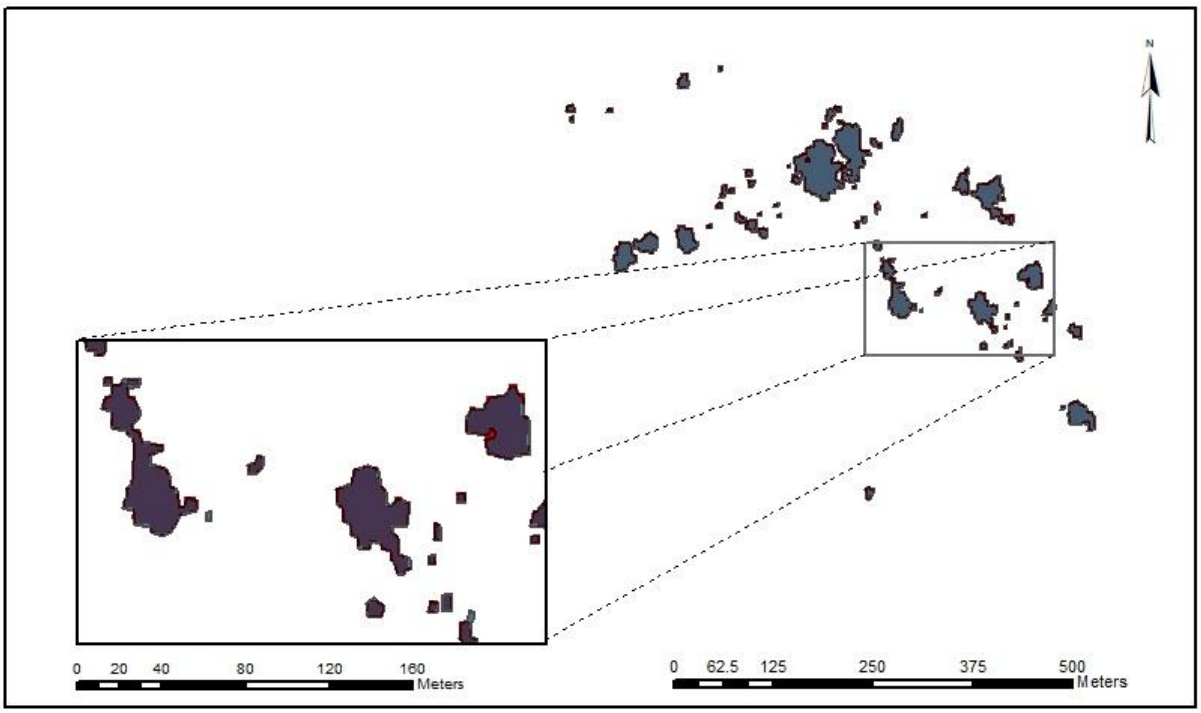


Figure 3-11: Common POIs - Raster to polygon

The reason is the comparatively extreme high density at some regions in individual tracks which causes a high density in the final raster even though it does not indicate a common stop and will result in discovery of fake common POIs (Figure 3-12).

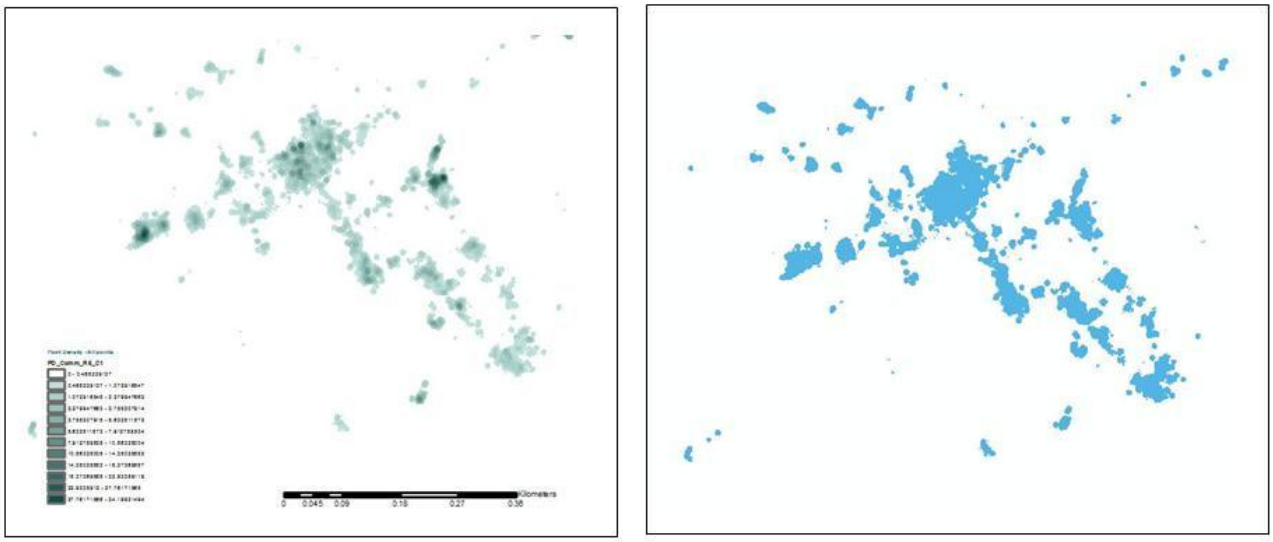


Figure 3-12: Point Density and produced raster from overlaying all points

3.3.4. Results

As it was illustrated in the previous section, the final result of this step contains polygons of 57 extracted common POIs with their corresponding tables of attributes, including id, geometry, coordinates of the centre (latitude, longitude), shape area and shape length.

3.4. Trajectory semantic annotation

3.4.1. Extracting related background information from online sources

Having extracted the location of points of interest it was possible to look for related information to each location from the suggested sources. This selection was based on the literature review on the semantic annotation of trajectories. Detailed reasoning is presented in the related sub-section.

Some of the platforms among online location based services were selected to be discussed and evaluated based on their relevance to the current study.

For this selection first of all the platform had to popular enough to have a rather complete dataset of places information generated by users. The second criterion was the available data and the type of information that could be extracted through their API (Application Programming Interface). Following are some of the candidate platform and a brief description of each:

Twitter[4]

Twitter allows its user to add location information to their tweets and these locations are accessible through the API.

Twitter converts lat,lon coordinates into actual places (in particular, neighbourhoods and cities). Street address, city, region and country code are some of the attributes that may be available in Geo Place Attributes.

Given latitude and longitude pair, an IP address, or a name, this request will return a list of all the valid places that can be used as the place_id. Through geo/search.Also through geo/reverse_geocode, given latitude and a longitude, it is possible to searches for up to 20 places that can be used as a place_id.

[4] http://www.twitter.com/

Flickr[5]

Flickr allows access through an API key which can be requested by application developers for non-commercial use.

Responds in rest (REST), xmlrpc (XML-RPC), soap (SOAP), soap2 (Alternate SOAP) json (JSON), *serial_php* (Serialized PHP) formats. Returns place_id with lat,long query and using place id it is possible to find information about the place.

The *flickr.places.find* method rounds up to the nearest place type to which place IDs apply. For example, if a street level address is passed, it will return the city that contains the address rather than the street, or building, itself. The information does not include type or name of the places.

Foursquare[6]

Foursquare as a straight places database, allows unauthenticated use, returns name and type of the places in a pre-defined radius around a specific latitude and longitude.

An array of categories is a response of, venue/categories, containing sub- and sub-sub- categories. Each top-level category contains an id, name, pluralName, icon, and categories (an array of child categories).

Returns a list of venues near the current location with the most people currently checked in.

Facebook[7]

Facebook Places contains location information and is based on *Factual API*. It's easily accessible for queries via *Graph API Explorer* by requesting access tokens. Or through *JavaScript Test Console.*

It is possible to retrieve information related to public places through the following request, narrowing down the search results by defining a center position (latitude, longitude) and a distance;

https://graph.facebook.com/search?q=coffee&type=place&center= 52.0122786063285,4.35937586720377 &distance=75&access_token=[access token retrieved through log in to Facebook]

Description of the place and description of geometry of the places are available through the places API in GeoJSON format as geometry type and coordinate arrays

Instagram[8]

As one of the most promising mobile applications of current time, Instagram was also considered as an option for the research purpose. People upload pictures along with geotags through this application.

But for its places service Instagram uses access to foursquare dataset.

GooglePlaces[9]

Nearby search in Google Places API provides information about places located in a specified distance from a location. The results can be refined by adding other parameters as place type and is provided through an API key or oAuth2 client ID, either in JSON or xml format. The response contains information about location, type, vicinity, photos (if available) and ratings. These results do not contain the distance parameter and are less detailed compared to Foursquare results.

After a general overview on the available platforms, Foursquare was selected as the primary source of information about the POIs. Some of the justifications for this choice are listed below:

- The main activity of Foursquare users is to do a "Check in" different places, therefore they are tend to update the database with information about the places that they are interested in. Hence the places database in continuously being updated with a large amount of location information from a significant number of this phone application users (around 25 million user at the time of documenting this research ("How Many People Use the Top Social Media?," 2013))

[5] http://www.flickr.com/
[6] http://foursquare.com/
[7] http://www.facebook.com/
[8] http://instagram.com/
[9] https://developers.google.com/places/

- The query response from Foursquare is the most complete compared to the others. The responses include many different fields from Name and Type (Category) of the place to the Address (street, city, country) and number of Check-ins and the Distance from the targeted location.
- It is free and the "venues platform" that provides the location information, does not require user authentication. It is it is possible have access to this database through oAuth 2.0 which is a standard used by most major API providers.

Venues platform is accessible through Foursquare's "API Explorer" and with forming a proper query it is possible to retrieve a list of places around one specific location. It is also possible to use scripted web applications such as 'apigee'[10].

The "Search" service is used from venues platform to locate all the places around a certain position. There are two options for performing this query. "Browse" and "Search". Both of these are acquiring latitude, longitude, a search radius and a limit for the number of responses per query. The difference is in the check-in criterion that is involved in the results of the "Search". Whereas "Browse" lists all the venues in the defined area (including personal places such as "my house", "Liny's place")

The Search option with a radius of 50 m was selected in this methodology since the number of check-ins indicates the interest of people in a place and that contributes to the logic of selecting common POIs and omitting personal preferences.

All the responses are structured as presented in Figure 3-13; whereas a part of response including one venue is shown in Figure 3-14.

```
{
    "meta": {
      "code": 200,
        ...errorType and errorDetail...
    },
    "notifications": {
        ...notifications...
    },
    "response": {
        ...results...
    }
}
```

Figure 3-13: Foursquare response structure

A corresponding place for each position was selected from these responses. This selection was based on two parameters of *distance* and *check-in*s. The main factor was considered as minimum distance from the centre of targeted polygons. The check-ins factor was considered where the difference in distance was less than 5 meters among number of results for the same place; in those cases the one with a larger number of check-ins was selected as the representative of matching polygon.

Not all the information from Fourquare response was saved as attributes of place, but among all, the most relevant attributes were selected to enrich the polygons with those attributes. These were Name, Category an Address for annotation purpose and distance and number of check-ins for indicating the precision and reliability of the provided details for purpose of future use and analysis (Figure 3-15). As the Foursquare data is completed by its users it can have imprecisions and incompleteness in different locations. This was evident in some of the responses where the closest suggested place to a polygon was located 25 meters away from its centre

[10] http://apigee.com/console/foursquare

Figure 3-14: Individual venue among response, from search

Place_Name	Location address	Place_Category	Distance	Check-ins Count	Location_lat	Location_lng
Tosti House	Markt 65	Sandwich Place	8	196	52.01171	4.35930
Grote Markt	Grote Markt	Plaza	9	1327	52.01165	4.35956
Bagels & Beans	Markt 61	Bagel Shop	14	376	52.01187	4.35920
Het Konings Huys	Markt 38 2611 GV	Bar	20	403	52.01162	4.35870
Willem van Oranje	Markt 48A	Diner	20	219	52.01174	4.35954
Cafe de Clipper	Markt 67	Bar	20	68	52.01151	4.35939
...	...	...	...	...	...	...

Figure 3-15: Information extraction from Foursquare

In addition to the information extracted from Foursquare other possible applications where also examined in order to fill the gaps of Foursquare dataset. For instance in some locations the closest response from Foursquare is located farther than 20 meters away from the centre of a small polygon, this indicates a gap in the dataset. To enhance the results Google Place was examined for some of these occurrences as the closest option to Foursquare. The access was made through its API explorer using a query structured as below:

https://maps.googleapis.com/maps/api/place/nearbysearch/json?location=52.0122786063285,4.359375 86720377&radius=75&sensor=false&key=[oAuth 2 client ID]

The experiments on 5 randomly selected common POIs, proved that Foursquare data was more complete compared to this Google Places option.

Facebook Places also provides similar type of information about places but its completeness is less than both Foursquare and Google Places.

3.4.2. Annotating the trajectories with the extracted information

The implementation of the annotation part was designed based on the computing platform and model for spatio-semantic trajectories to associate the trajectories with geographic metadata, presented by Yan et al. (2010),(2011). The main idea is to enrich the basic abstraction of trajectories (result of previous two steps) to a higher-level abstraction (e.g. shop, church). The concepts about level of granularity in providing semantic information and the type of intersections applied on the data were mainly derived from these works.

Topological intersection was chosen over the geometrical distance (calculating distance of each point from points or lines of interest, commonly used for POIs and road networks) in this dissertation. The reason is that the common POIs were treated as regions and were extracted as polygons.
To do so the polygons were first enriched by their corresponding extracted information. The polygons' table was updated by information extracted from foursquare (Figure 3-16).
A spatial join was then applied between the derived polygons of common POIs and the recorded points of stops for each trajectory. Therefore all these points were assigned with the attributes of the containing polygons (including semantic information of each place).

OBJECTID	...	center_lng	center_lat	Name	Category	Address	Distance	Checkins
...	...	...	...	...	...	...	...	...
49		4.3622229	52.0101031	HEMA	Department Stores	Molslaan 33	8	838
50		4.3639142	52.0098823	Bruna	Bookstore	Gasthuislaan 66	14	105
51		4.3627445	52.0097307	Dixons Paradijspoort	Electronics Store	Paradijspoort 32	5	65
...	...	...	...	...	...	...	...	...

Figure 3-16: Attributes added to common POIs from Foursquare

In the final step the enriched extracted points of stops were again joined with the initial trajectory. The stops which are not corresponding to any of the common POIs are tagged as *Unknown stops*. Figure 3-17 illustrates the followed approach. This process was repeated for each trajectory that had corresponding layer of extracted stops.

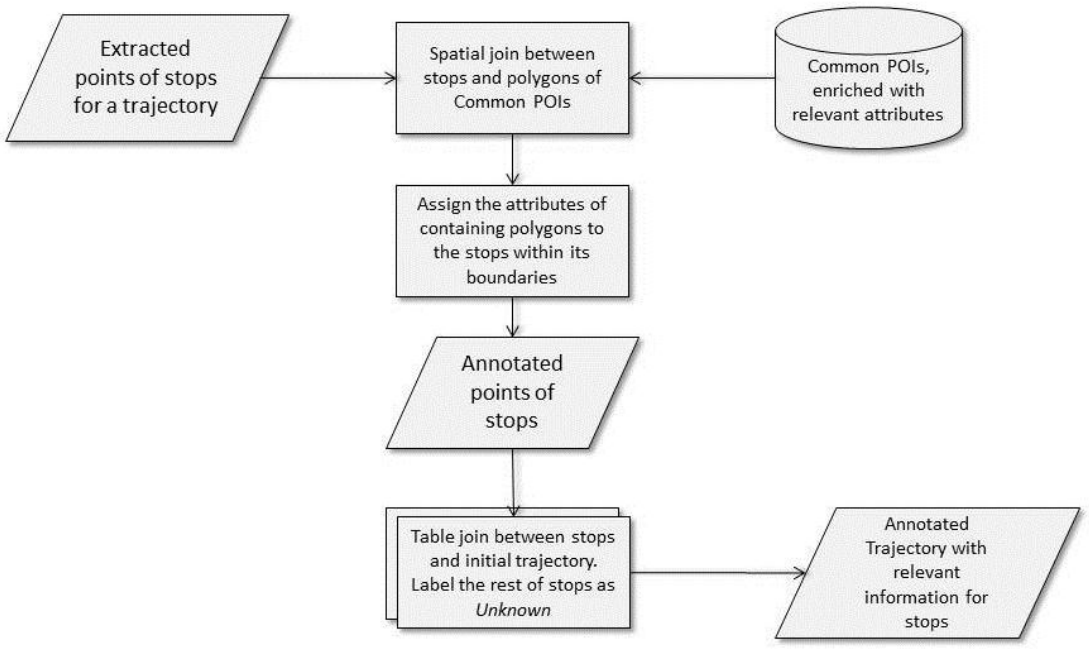


Figure 3-17: Trajectory semantic annotation flow

3.4.3. Results

The final results for 2 randomly selected trajectories are presented in Figure 3-17.

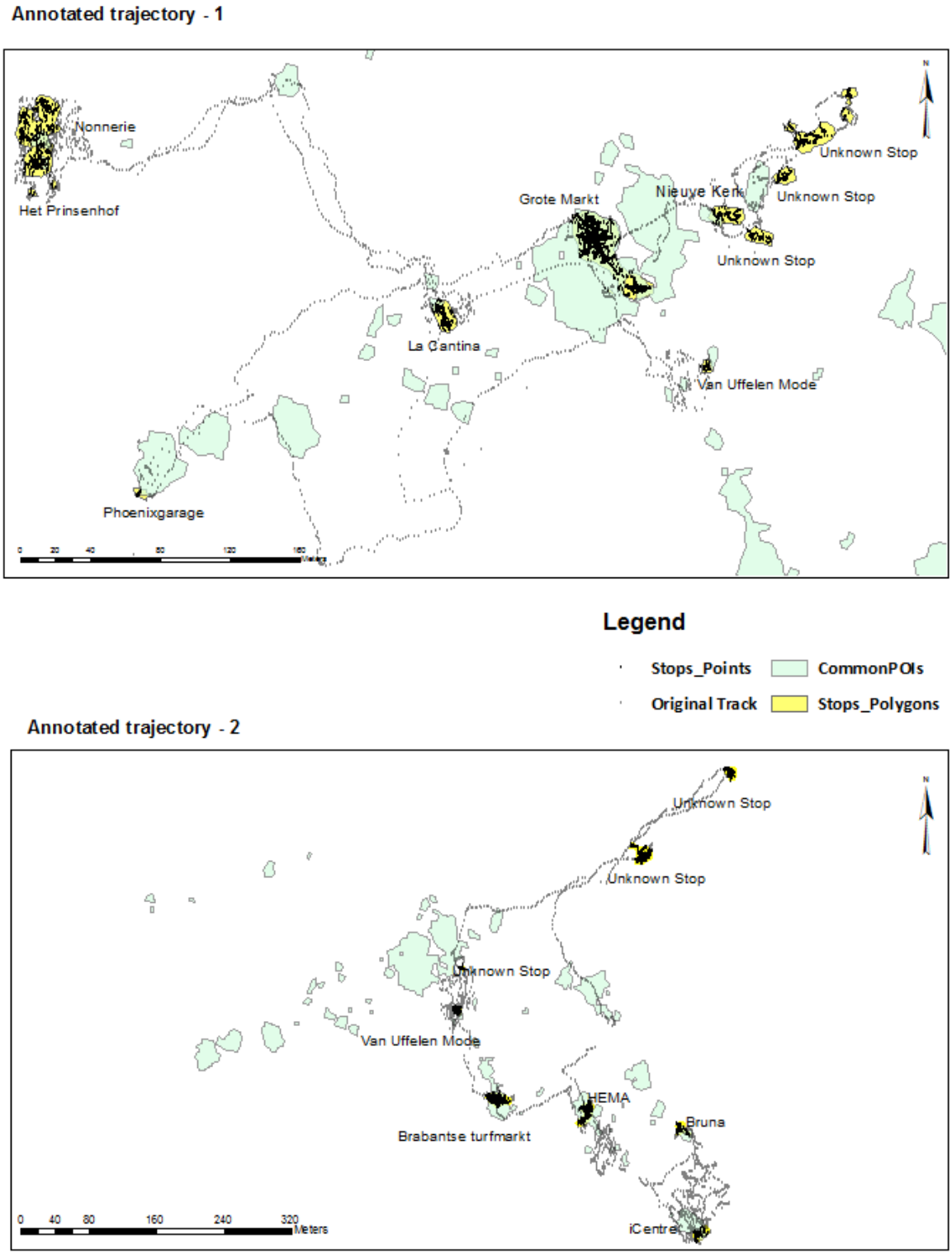


Figure 3-18: Trajectory annotation examples

All three main steps for an individual trajectory, from discovering individual stops to annotating them are illustrated for another single track in Figure 3-18.

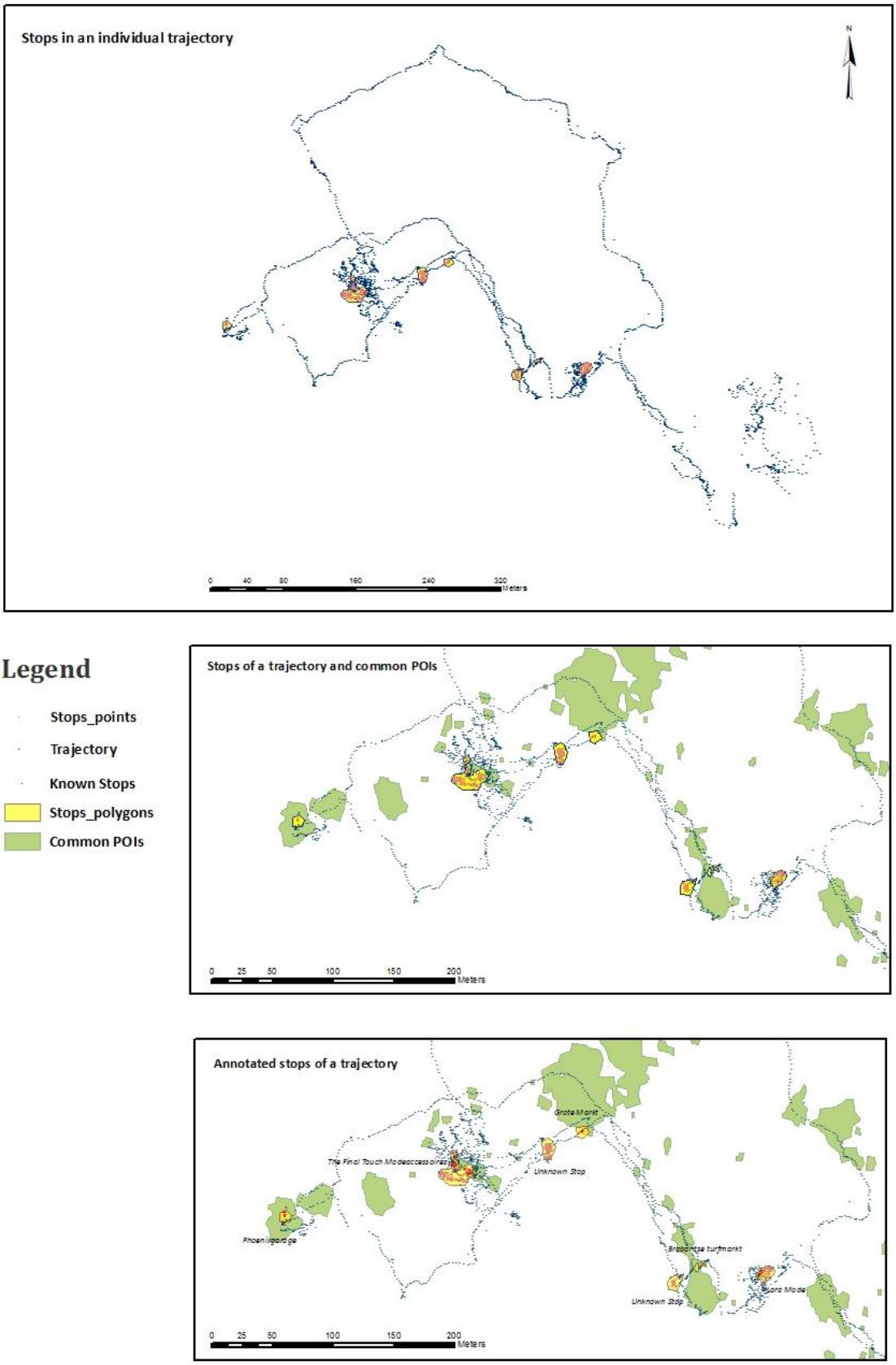


Figure 3-19: Trajectory annotation example_ workflow illustration

3.5. Validation by visual interpretation an map overlay

In order to validate the results obtained from Fourquare and added to the trajectories, a map overlay was performed to inspect and verify the geographic position of extracted common POIs.

This validation was to check the reliability of the discovered polygons of significant places and not to validate the extraction of stops in individual trajectories. Individual stops can be validated based on the information provided by the GPS holder about his movements. This type of information may be provided along with the movement dataset, as a list of places that the moving object has stopped. Unfortunately this data was not included in the available dataset used for the current research. Therefore, the common POIs (which are resulted from the individual stops) where validated to check the level of accuracy, relevance and reliability of the discovered places with respect to the background geography and the nature of the analysed dataset.

For this purpose OpenStreetMap[11], a well-known example of VGI source was used to compare the obtained results with the existing information in this source.

In addition to OSM (OpenStreetMap), where data was not available about a specific position Google Maps[12] was also used occasionally to check the reliability of OSM data.

Figure 3-20 represents the results for this validation. The polygons which are shown as 'Correct' are those which the type of place, name of place, and the distance compared to the size of polygon are logical and in consonant with OSM. There are 7 polygons which are labelled as 'Correct_Less certain' the reason for not indicating them as correct is different with respect to the polygon. In some cases the position is corresponding with the features in OSM but data from Foursquare indicates a larger distance (Figure 3-21, row 25, 28). Some are chosen based on distance and check-ins but compared to the OSM and the context, the first result (without considering check-ins) seems more logical (Figure 3-21, row 38). Some are assigned attributes from Foursquare where the distance and check-ins seem to be correct and also it does not look irrelevant on OSM. But taking into account the nature of movement and the features present in the target position it can be interfered differently. For example in Figure 3-21, row 16, the polygon is assigned with attributes of a cocktail bar where in reality it seems the location was crowded as it is located in a junction were a main road is dividing into two branches.

There are 6 features which the closest location found in Foursquare is located farther than 16 meters away from the inquired coordinates, these have less certainty due to incompleteness of Foursquare's location database at that position.

Generally were the results are less accurate the reasons can be grouped as following:

- Accuracy of GPS which is about 10-15 meters,
- Narrow streets in the studied area
- High amount of activities in a small area
- Large number of small potential POIs placed next to each other in a crowded city centre.
- Accuracy of Open Street Map which is about 6 meters (Haklay, 2010).

Where the reliability of extracted information is low due to above mentioned reasons, it is possible to annotate the polygons with a lower degree of granularity. For example annotation can be performed only based on the address and type of place, instead of including name of the place as well.

[11] http://www.openstreetmap.org/
[12] https://maps.google.nl/

Figure 3-20: Validation, Map overlay

Center Coordinates	Area	Place_Name	Place_Address	Place_Category	Distance	Check-ins
52.0127800318308,4.35736801485963	21.01016942	Trekpleister	Choorstraat 8	Cosmetics Shop	4	36
52.0126298016899,4.3566900067127	191.5916886	Huyser	Choorstraat 1	Bookstore	15	339
52.0123459483553,4.35951584692912	46.07239937	Cafe Restaurant 't Vermeertje	Markt 58	Restaurant	16	63
52.0122954413536,4.35535803266076	27.29797717	Het Prinsenhof	Sint Agathaplein 1	History Museum	23	324
52.0123056542911,4.35464649176049	83.27110225	Nonnerie	Sint Agathaplein	Café	12	39
52.0122786063285,4.35937586720377	104.2286446	Rijwielsporthuis Piet Vonk	Voldersgracht 20	Bike Shop	3	29
52.0121915106296,4.35467678279761	25.70299817	Het Prinsenhof	Sint Agathaplein 1	History Museum	23	324
52.0121540644836,4.35928769601338	26.09254177	Cafe Restaurant 't Vermeertje	Markt 58	Restaurant	11	63
52.0121188008422,4.36060641290805	258.2299302	Nieuwe Kerk	Kerkstraat	Church	15	517
52.0119667390748,4.36021757331093	91.88212269	Nieuwe Kerk	Kerkstraat	Church	16	517
52.0116982357278,4.35863539144728	9.260948468	De Liefhebber	markt 20	Café	7	38
52.0116082483011,4.3579092045852	61.58580639	De Koornbeurs	Voldersgracht 1	Nightclub	4	205
52.0115453500958,4.35877695419183	94.16790089	Stadhuis	Markt 87	City Hall	5	363
52.011490599275,4.35795286646548	53.11468288	La Cantina	Markt 3	Mexican Restaurant	12	236
52.0114660810785,4.35984259562417	9.260994727	Döner King	Oude Langendijk 19	Falafel Restaurant	5	141
52.011528632293,4.36180933745151	354.0877286	Breeze Cocktailbar	Burgwal 33	Cocktail Bar	8	177
52.011410004887,4.3576018988626	39.03293791	Steendam Herenmode	Wijnhaven 20-21	Boutique	10	57
52.0113674165339,4.36254707533352	17.80516822	Popocatepetl	Beestenmarkt 35	Mexican Restaurant	5	196
52.0113997431322,4.35745785650011	88.01042601	De Wijnhaven	Wijnhaven 22	Bar	12	757
52.0117460391684,4.35941949880082	4704.697274	Grote Markt	Grote Markt	Plaza	9	1327
52.0112496432092,4.35842269466987	24.85071699	t Boterhuis	Markt 15-17a	Restaurant	8	2396
52.0112387443072,4.35738149869363	52.80803853	Halfords	Wijnhaven 17	Bike Shop	10	47
52.0112418126823,4.3602587092188	53.53598079	Van Uffelen Mode	Jacob Gerritstraat 12	Clothing Store	25	30
52.0111526723017,4.35811884735233	12.34808135	De Waag	Markt 11	Café	15	1006
52.0111621304542,4.36113406334051	27.78316695	De Beierd	Burgwal 18	Gastropub	21	353
52.0111294568368,4.35847868705889	12.34808725	Atelier Art of Cut	Markt 25	Salon / Barbershop	1	24
52.0113780800899,4.36233086176957	1139.752873	Beestenmarkt	Beestenmarkt	Plaza	9	1245
52.0111314721039,4.36270585983859	74.81748704	Barrique	Beestenmarkt 33	Wine Bar	22	184
52.0111013534354,4.36002024524504	32.89540807	Van Uffelen Mode	Jacob Gerritstraat 12	Clothing Store	3	30
52.0110948631312,4.35776603396421	100.6161741	The Final Touch Modeaccessoires	Wijnhaven 14	Jewelry Store	18	5
52.0110438118447,4.35989531682616	30.5342021	De Tuinen	Jacob Gerritstraat 10	Drugstore / Pharmacy	5	48
52.0110017717797,4.35720183648712	18.522186	Plan B	Boterbrug 13-15	Ice Cream Shop	10	302
52.0110194238018,4.35797880249837	166.8215759	Umai	Wijnhaven 11	Sushi Restaurant	15	354
52.0109454122667,4.35818450043821	111.8177831	Coffee Company	Markt 19-21	Coffee Shop	9	1248
52.010814982793,4.36028305875744	75.91750053	Telfort	Jacob Gerritstraat 4	Electronics Store	1	10
52.0108534930184,4.35678607273857	614.4210046	Miles	Oude Delft 125	Mediterranean Restaurant	20	120
52.0107988566139,4.35606518856015	506.5537906	DSB		Fraternity House	11	2
52.0106535398056,4.36037074745161	12.34821467	Delftse Markt	Burgwal	Mall	10	521
52.0106518567309,4.3556568555137	720.076511	Phoenixgarage	Phoenixstraat 29	Parking	30	427
52.0104934391369,4.3631080589079	665.8724276	ANWB Winkel	Pynepoort 3	Miscellaneous Shop	30	42
52.0103038113208,4.36140168132439	40.3096159	Lara Mode	Molslaan	Women's Store	5	19
52.0101776227742,4.36283158673533	12.3483414	Vero Moda	Paradijspoort 7-11	Women's Store	21	35
52.0100847609146,4.36108462299856	9.261277015	nobel	Molslaan	Shoe Store	5	1
52.0101329280526,4.3634459880038	201.1599017	Blokker	Kruisstraat 44	Miscellaneous Shop	12	118
52.0102824716581,4.36063694322479	1125.75658	Brabantse turfmarkt	Brabantse Turfmarkt	Market	2	4
52.0100311823971,4.36266751892972	21.72590884	Luc Snackkiosk	Paradijspoort	Food Truck	2	39
52.0100035067872,4.3633481308303	15.43548535	Blokker	Kruisstraat 44	Miscellaneous Shop	27	118
52.0099106449275,4.36263425528327	12.34841475	Luc Snackkiosk	Paradijspoort	Food Truck	11	39
52.0101031143132,4.36222291784097	919.3746628	HEMA	Molslaan 33	Department Stores	8	838
52.0098823168717,4.36391415021063	210.500558	Bruna	Gasthuislaan 66	Bookstore	14	105
52.0097307250741,4.36274452874179	30.87115947	Dixons Paradijspoort	Paradijspoort 32	Electronics Store	5	65
52.0097086123292,4.36264559425536	21.32312416	We Men	Paradijspoort 26-30	Men's Store	5	15
52.0096996050057,4.36223952307477	56.16092663	Jack & Jones Delft	Paradijspoort 22	Clothing Store	18	107
52.0096018112757,4.36288532060089	83.16968066	Lucardi	Paradijspoort 58	Jewelry Store	7	22
52.0090163449312,4.36369378886481	20.51630351	ICI Paris XL	Bastiaansplein 2	Cosmetics Shop	9	42
52.0089348522834,4.36403478988715	733.0501562	iCentre	Bastiaansplein 5	Electronics Store	3	532
52.0080305617156,4.36018894190267	112.3303247	Lazuli	Lange Geer 72	Arts & Crafts Store	7	4

Figure 3-21: Annotated polygons of common POIs

3.6. Workflow design

The final workflow diagram is presented in this section. This flow diagram summarizes all the steps, inputs, procedures, determined parameters and outputs and the connection between these steps.

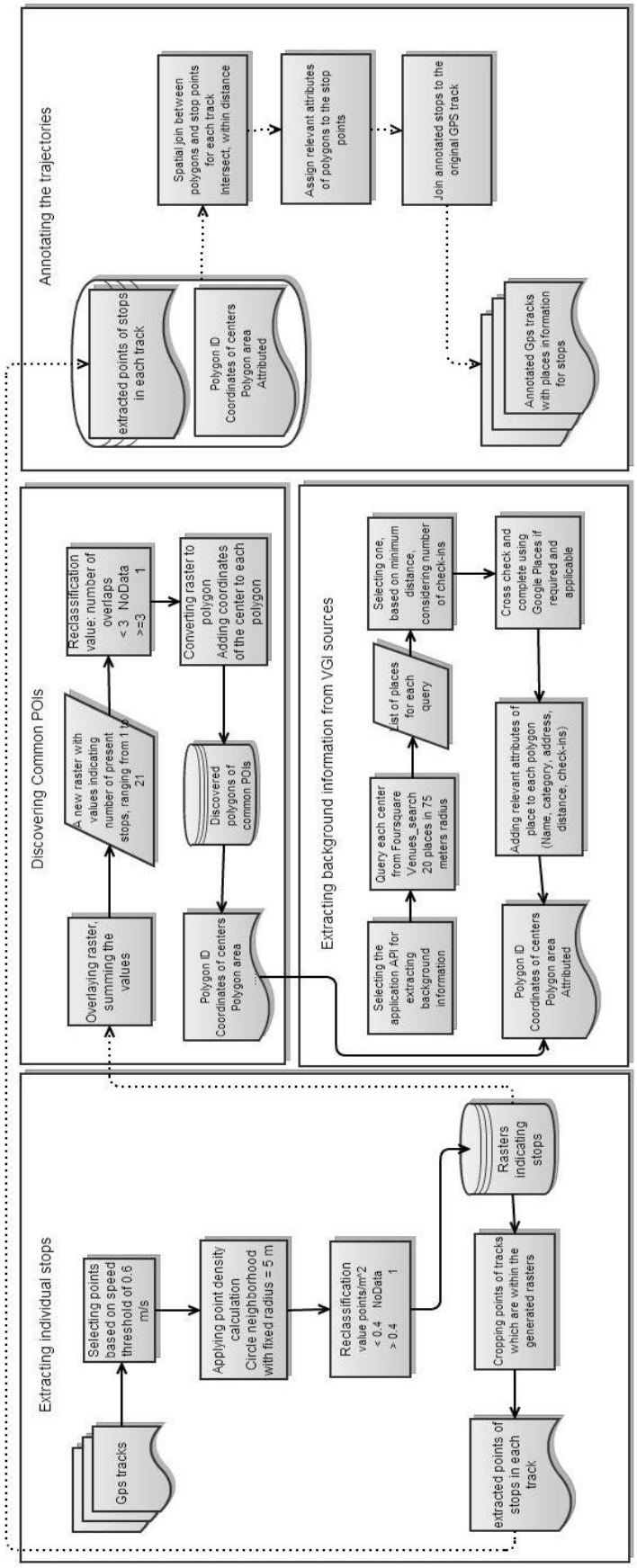


Figure 3-22. Final workflow design

4. DISCUSSIONS AND RECOMMENDATIONS

4.1. Discussions

4.1.1. Parameter selection for extracting stops

The main fixed parameters during this experiment were *speed threshold*, *neighbourhood radius*, *minimum density* for a place to be considered as a stop, and *minimum number of overlaps* for rasters of stops to form a common POI. Each of these values can be altered and the results may have slight or significant changes according to the change in these values. Defining a large radius (e.g. 15 m) for the point density calculation will result in the production of rather large areas with over all lower densities (smaller maximum density compared to the r = 5 m results). Therefore the minimum value for density threshold should be also lower to avoid omitting many parts from the results. But doing so will cause a lower accuracy as the remaining regions are big areas with low density and the determined places are not focused around the present points of the trajectory. On the other hand an extremely small radius will result in shaping dense discrete small regions that are not reliable because of their size compared to GPS device's accuracy and the real world parcel sizes.

The speed limit can be selected lower than 0.6 m/s if the target is to only determine real physical stops; however selection of a relatively high threshold (0.6 m/s) can have advantages since it will result in detection of regions that people are tend to *slow down* and are wandering around. A clear example in the studied dataset could be the open market area of Delft that is extracted as a large polygon (see Figure 4-1). This can be due to these types of slow movements around a *region of interest* instead of actually stopping and staying in a *point of interest* like a restaurant for a certain time interval.

Despite the similarities of followed approach with CB-SMoT algorithm, one of the differences is in including the "duration of stay" parameter. A. T. Palma et al. (2008) imported the candidate stops are imported to a model and the generated clusters were checked against a predefined list of places, when the cluster matches a candidate stop and the duration of stay (the time interval between the first and last point in the cluster) is higher than a certain threshold that cluster is labelled as *Stop*. On the contrary the present research is aiming for defining the potential candidate stops, therefore the duration of stay was not applicable in a form that calculates the time interval between points. This temporal aspect was included in the selection of minimum density value assuming that at least 1 minute is required for a moving object to stay in a point density calculation neighbourhood to generate a stop. Including the time as duration of stay would have required a spatio-temporal clustering.

The density value indicating number of points per square meter, could have been set to a higher threshold for more certain results with more reliability of presence in a certain position. Yet it would have limited the discovery of *potential stops* and could have possibly affected the discovery of common POIs as well. The value of 0.4 points/m^2 was selected based on an assumption of 30 points ideally being recorder by the GPS device in 1 minute. In reality the GPS signal may be lost and the number of recorded points might be less than expected, or some points might be removed during pre-processing stage. Therefore, the number of recorded points is usually less than 30 in 1 minute.

In this setting for the fixed values the reliability might be lower for the annotation but most of the potential places are discovered with the followed procedure.

The last parameter was the number of overlays indicating the significance of a place. That again a higher value will result in more certainty about interests of more people in that place but less potential positions to be discovered.

4.1.2. Generated raster for common POIs

After generating the raster of overlaying all stops and before converting it to a polygon some post classification filters were applied to refine the results, omit the noises and separate polygons that are connected with a narrow link between them. As a result of boundary clean filter applied on the output raster, some regions with small areas were treated as noise. Omitting these regions would not have had a negative impact on the final results because the sparse and small regions are not indicators of a dense area. This is because of the point density calculation results that the value of density always reduces from the centre to edges.

4.1.3. Size of extracted polygons

As it is evident from Figure 3-11 there exist large polygons in the produced results. The area of generated polygons varies between 9 and 4750 m^2. After the map overlay it can be inferred that this type of behavior occurs in the places that are genuinely crowded as an area of interest. For instance the largest polygon is overlapping on the market area of the city centre as people tend to walk around slowly and spend time in different positions of the same area. Where this is not an incorrect result it prevents extraction of possible smaller polygons in the same location.

4.1.4. Including additional characteristics to the analysis

It is also possible to implement the suggested method for a group of trajectories among the whole dataset based on specific criteria. This criteria can be either temporal or it can be relevant to the characteristics of each moving entity. As an example the input trajectories can be selected based on occupation or activity type of the GPS holder. This can produce results for questions like: *where are the common POIs for tourists?* Or *where do people go for shopping the most?*

The temporal aspect can be used when selecting a group of trajectories based on time interval, for example daily activities can be studies and compared if the input data is divided into four different groups for four days. Therefore it will be possible to have observations like: *which places do people go on Saturday?*

4.2. Recommendations about points of discussion

- Parameters and reliability of results

The presented model in current workflow is set with the parameters that will discover optimum number of potential POIs and generate a base layer of significant places used for annotations. But the parameters for extracting individual stops were rather high; hence lower values can be set for more certain and accurate results on individual tracks.

- Selection of places from VGI

For a more sophisticated level of analysis in order to choose the best match from Foursquare, the distance and ranking units can be unified and a weight can be assigned to each parameter. For example if the distance has a weight of 0.8 and the rankings and checking have 0.2 of importance a new influence factor can be generated and selection will be made based of this factor.

The issue of assigning the closes or more related POI from candidate POIs has been taken into consideration in (Xie et al., 2009), where they takes this to a higher level and consider the duration of stay as an indicator of activity and assigns the place and activity type to the stops based on that. For example if a university and a shop are located in a close distance from each other but the moving object stops for hours in that location this is more likely that he was in the university than in the shop.

- In the generated raster for extracting common POIs 21 values are produced that each of them represents a degree of significance for its matching region. If the reclassification contains more than 1 value it is possible to talk about degree of importance or reliability of significance of a place. For example the classification would include 3 intervals of 1-6, 7-15 and 15-21 and a separate value could

be assigned to each of these intervals. Then it is possible to say the third interval is of the most importance and it is more accurate compared to the others as well.

- Polygons can be divided into smaller areas; it could have been done using parallel lines with a defined distance, or a grid with a certain cell size to cut the polygon into smaller polygons. The challenge would be to select a proper measure for selecting the grid or parallel lines. Two main reasons for the possible complications are:
    1. If parallel lines are used, the orientation of them will be affecting the division. For example using horizontal lines for dividing a horizontally wide polygon will not enhance the results on horizontal edges of the initial polygon.
    2. The area for the extracted polygons which varies significantly between 9 and 4700 and makes the selection of grid size challenging.

5. CONCLUSION AND FUTURE WORK

5.1. Conclusion

The availability of massive amount of movement data collected through different location acquisition technologies, and the considerable increase of this availability in different fields suggests the necessity of providing structures and tools for more efficient analysis of these data in various application domains.
Current thesis reports on exploring an analytical methodology used for developing the workflow (as the main objective of the research) and generating results with the best fitting parameters according to the application purpose and data characteristics.
Each trajectory was first divided to sub sequences of points indicating stops and moves. SMoT (Alvares et al., 2007) and similar algorithm that check the presence of the moving object against the pre-defined places of Interest, clustering based SMoT (Andrey Tietbohl Palma et al., 2008) and other enhancements to this algorithm like Intersection Based SMoT or Direction Based SMoT (Rocha et al., 2010), and also detecting suspensions of patterns by Orellana et al. (2012) were among the alternatives for discovering important parts of a trajectory.
Extracting of the stops was performed based on a point density calculation and the corresponding parameters were selected based on characteristics of the movement. It was also considered that these stops will further be used as inputs for discovering common POIs. The resulted stops include all the positions for a possible different behaviour of the pedestrian during the trip. This was due to the selection of parameters which were set to the minimum requirements for defining a stop. Different methods were studied and examined and at the end the indicators of a stop were set as maximum instant speed of 0.6 m/s and 0.4 points/m^2 for the point density.
Using the individual stops, common POIs were discovered based on raster overlay of extracted stops for all the trajectories in the dataset. This produced polygons of significant places according to the common interests in movement behaviour of studied dataset. The results were satisfactory for the less dense regions but for the main crowded centres such as the open market area the overlay produced a large polygon covering a considerable area. This was a problem when annotating all the stops in this large polygon with the same attribute (the attributes of polygon were assigned based on the closest place to the centre of polygon). Results can be enhanced with human interaction by dividing the large polygons into smaller parcels for a more detailed level of annotation, or a more generic attribute can be assigned to the polygon which will increase the reliability of this attribute but the level of details in the annotation will decrease.
The corresponding data to the common POIs was extracted from Foursquare and was added in 5 different groups of attributes to the stops, including *name of the place*, *type of the place (category), address, distance from the centre of polygon* and *number of check-in*s. These attributes were selected based on their contribution to aim of this research and similar studies.
All the online web or mobile applications that provide the facility of uploading geolocated information for their users, could be a candidate for the aim of conducted research. Since these applications have a geospatial data repository which is continuously being updated and completed by user contribution. Depending on the selection of application and the data that is aimed to be retrieved, it is possible to request responses for a location dependant query through the API of these services. As each of the platforms and applications provide a different level of details and different type of information about a queried geographic coordinate it is possible to use an alternative to complete the gaps of one selected platform when necessary (e.g. occasionally using Google Places to fill the gaps of Fourquare).

Yet a connection between online data repositories (VGI sources) and the designed workflow is needed for a fully automated approach. Applying a function for automatic selection of the most relevant feature among the responses for location query will also reduce the required human interaction.
The last step was to integrate the extracted information from Foursquare with the determined stops on each trajectory. Map matching algorithms and applying spatial joins between trajectories and given set of regions (Yan et al., 2011) were some of the options considered for this purpose. The annotation was done based on the spatial join of polygons for places and points for the trajectories. Where the points were within the boundaries and in a certain distance from a place's polygon they were annotated with the corresponding information about that place.
The final workflow was developed in a way that it requires minimum contribution from the user's side and annotates the trajectory stops with the relevant information of places extracted from the dataset. While it was also considered that parameters can be imported from the user to the model for more possible analysis
The designed workflow is generic for different types of human movements in the city, since even with the use of different transportations the stops are still going to be discovered with the proposed method. Whereas for a completely different application domain (e.g. animal movements) this workflow will not necessarily generate correct results, as the movement type, spatio-temporal characteristics, definition of stops and moves, and the type of 3rd party sources for background information are thoroughly different from those studied in the current research.
At the end the integration of spatial, non-spatial and trajectory data produced tracks that are annotated with the relevant attributes of places in which the moving object showed a different behaviour compared to the rest of trip. The data was structured and enriched for further place-related analysis of movement with corresponding attributes of significant places extracted through a holistic view on the dataset.
The possible follow ups to this work are presented in the next sub-section.

5.2. Future work

In this research the temporal aspect was not directly taken into account. The workflow can also be applied for a selection of trajectories in a certain time intervals. If the temporal aspect of these movements also gets included in the process, much more interesting results can be achieved. Studying periodic behaviours (Li et al., 2010) or discovering frequent activities (Bamis et al., 2010), event detections, discovering similar semantic behaviours can be mentioned in relation to this type of research.

Another possibility is to extract the semantic information from Foursquare with respect to the *time* component as well. The foursquare venues' API provides an option of queries that retrieve results based on the time of the day and the most possibility of a check-in to a place at that time of the day. This type of queries will prevent receiving irrelevant but close places. For instance a city centre it is quite likely to have bars and cafes in the same neighbourhood and a close distance from each other, but in a morning time a bar will not be a suggested venue compared to a café.
From the implementation aspect, next step would be to develop an application, based on the suggested workflow. All the tools and followed procedure were selected and designed with a consideration of future implementations for a fully automated flow. This flow can be developed further by making a connection between all the proposed steps in a proper environment to have a homogenous and effective standalone application with minimum user assistance.
Consequently, all the followed steps will be placed in a well-structured flow to develop an application that receives the movement data as GPS tracks and delivers the annotated data as the output. In this type of structure parameters can also be received from the user rather than being fixed and this will produce an interactive workflow.

LIST OF REFERENCES

Agrawal, R., Gehrke, J., Gunopulos, D., & Raghavan, P. (1998). Automatic subspace clustering of high dimensional data for data mining applications. *SIGMOD Rec., 27*(2), 94-105.

Alvares, L. O., Bogorny, V., Kuijpers, B., De Macedo, J. A. F., Moelans, B., & Vaisman, A. (2007). *A model for enriching trajectories with semantic geographical information*.

Andrienko, G., Andrienko, N., Bak, P., Keim, D., Kisilevich, S., & Wrobel, S. (2011a). A conceptual framework and taxonomy of techniques for analyzing movement. *Journal of Visual Languages and Computing, 22*(3), 213-232.

Andrienko, G., Andrienko, N., Hurter, C., Rinzivillo, S., & Wrobel, S. (2011b). *From movement tracks through events to places: Extracting and characterizing significant places from mobility data*.

Andrienko, G., Andrienko, N., & Wrobel, S. (2007). Visual analytics tools for analysis of movement data. *SIGKDD Explor. Newsl., 9*(2), 38-46.

Ankerst, M., Breunig, M. M., Kriegel, H. P., & Sander, J. (1999). OPTICS: Ordering Points to Identify the Clustering Structure. *SIGMOD Record (ACM Special Interest Group on Management of Data), 28*(2), 49-60.

Bamis, A., & Savvides, A. (2010). *Lightweight Extraction of Frequent Spatio-Temporal Activities from GPS Traces*. Paper presented at the Proceedings of the 2010 31st IEEE Real-Time Systems Symposium.

Bogorny, V., Avancini, H., de Paula, B. C., Kuplich, C. R., & Alvares, L. O. (2011). Weka-STPM: A Software Architecture and Prototype for Semantic Trajectory Data Mining and Visualization. *Transactions in GIS, 15*(2), 227-248.

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

Ester, M., Kriegel, H., Sander, J., & Xu, X. (1996). *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*. Paper presented at the Second International Conference on Knowledge Discovery and Data Mining.

GeoPKDD. (2013). Retrieved February 2013, from http://www.geopkdd.eu./

Giannotti, F., Nanni, M., Pedreschi, D., Pinelli, F., Renso, C., Rinzivillo, S., & Trasarti, R. (2011). Unveiling the complexity of human mobility by querying and mining massive trajectory data. *The VLDB Journal, 20*(5), 695-719. doi: 10.1007/s00778-011-0244-8

Giannotti, F., & Pedreschi, D. (2008). *Mobility, data mining and privacy: geographic knowledge discovery*. DE: Springer Verlag.

Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. *GeoJournal, 69*(4), 211-221.

Hägerstrand, T. (1970). What about people in Regional Science? *Papers of the Regional Science Association, 24*(1), 6-21.

Haklay, M. (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B: Planning and Design, 37*(4), 682-703.

Heipke, C. (2010). Crowdsourcing geospatial data. *ISPRS Journal of Photogrammetry and Remote Sensing, 65*(6), 550-557.

Hong, L., Ahmed, A., Gurumurthy, S., Smola, A. J., & Tsioutsiouliklis, K. (2012). *Discovering geographical topics in the twitter stream*. Paper presented at the Proceedings of the 21st international conference on World Wide Web, Lyon, France.

How Many People Use the Top Social Media? (2013, 03/02/2013). Retrieved 03/02/2013, 2013, from http://expandedramblings.com/index.php/resource-how-many-people-use-the-top-social-media/

Idrissov, A., & Nascimento, M. A. (2010). *A Trajectory Cleaning Framework for Trajectory Clustering*.

Jain, A. K., & Dubes, R. C. (1988). *Algorithms for clustering data*: Prentice-Hall, Inc.

Kalnis, P., Mamoulis, N., & Bakiras, S. (2005). *On discovering moving clusters in spatio-temporal data*.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: an introduction to cluster analysis*: Wiley.

Ketterlin, A. (1997). Clustering Sequences of Complex Objects *KDD* (pp. 215-218).

Kisilevich, S., Krstajic, M., Keim, D., Andrienko, N., & Andrienko, G. (2010a). *Event-based analysis of people's activities and behavior using Flickr and Panoramio geotagged photo collections*.

Kisilevich, S., Mansmann, F., & Keim, D. (2010b). *P-DBSCAN: a density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos*. Paper presented at the Proceedings of the 1st International Conference and Exhibition on Computing for Geospatial Research; Application, Washington, D.C.

Knoblauch, R., Pietrucha, M., & Nitzburg, M. (1996). Field Studies of Pedestrian Walking Speed and {Start-Up} Time. *Transportation Research Record, 1538*(-1), 27-38.

Laube, P., Imfeld, S., & Weibel, R. (2005). Discovering relative motion patterns in groups of moving point objects. *International Journal of Geographical Information Science, 19*(6), 639-668.

Lee, J.-G., Han, J., & Whang, K.-Y. (2007). *Trajectory clustering: a partition-and-group framework*. Paper presented at the Proceedings of the 2007 ACM SIGMOD international conference on Management of data, Beijing, China.

Li, Z., Ding, B., Han, J., Kays, R., & Nye, P. (2010). *Mining periodic behaviors for moving objects*. Paper presented at the Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, Washington, DC, USA.

MODAP. (2013). Retrieved February, 2013, from http://modap.org/

Move-Cost. (2012). Retrieved Nov 21, 2012, from http://www.move-cost.info/

Naaman, M. (2011). Geographic information from georeferenced social media data. *SIGSPATIAL Special, 3*(2), 54-61.

Nanni, M. (2002). Clustering methods for spatio-temporal data. *Pisa, Italy: University of Pisa*.

Ng, R. T., & Han, J. (1994). *Efficient and Effective Clustering Methods for Spatial Data Mining*. Paper presented at the Proceedings of the 20th International Conference on Very Large Data Bases.

Orellana, D., Bregt, A. K., Ligtenberg, A., & Wachowicz, M. (2012). Exploring visitor movement patterns in natural recreational areas. *Tourism Management, 33*(3), 672-682.

Palma, A. T., Bogorny, V., Kuijpers, B., & Alvares, L. O. (2008). *A clustering-based approach for discovering interesting places in trajectories*. Paper presented at the Proceedings of the 2008 ACM symposium on Applied computing, Fortaleza, Ceara, Brazil.

Palma, A. T., Bogorny, V., Kuijpers, B., Alvares, L. O., & Acm. (2008). *A Clustering-based Approach for Discovering Interesting Places in Trajectories*. New York: Assoc Computing Machinery.

Rocha, J. A. M. R., Oliveira, G., Alvares, L. O., Bogorny, V., & Times, V. C. (2010, 7-9 July 2010). *DB-SMoT: A direction-based spatio-temporal clustering method*. Paper presented at the Intelligent Systems (IS), 2010 5th IEEE International Conference.

Sander, J., Ester, M., Kriegel, H.-P., & Xu, X. (1998). Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications. *Data Min. Knowl. Discov., 2*(2), 169-194.

Sheikholeslami, G., Chatterjee, S., & Zhang, A. (1998). *WaveCluster: A Multi-Resolution Clustering Approach for Very Large Spatial Databases*. Paper presented at the Proceedings of the 24rd International Conference on Very Large Data Bases.

Spaccapietra, S., & Parent, C. (2011). *Adding meaning to your steps*. Paper presented at the Proceedings of the 30th international conference on Conceptual modeling, Brussels, Belgium.

Spaccapietra, S., Parent, C., Damiani, M. L., de Macedo, J. A., Porto, F., & Vangenot, C. (2008). A conceptual view on trajectories. *Data & Knowledge Engineering, 65*(1), 126-146.

Spatial Metro. (2012, 2012). Retrieved Nov, 2012, from http://www.bk.tudelft.nl/onderzoek/onderzoeksprojecten/spatial-metro/

Trochim, W. M. K. (2006, 10/20/2006). Research Methods Knowledge Base Retrieved Aug 22, 2012, from http://www.socialresearchmethods.net/kb/contents.php

Van Langelaar, C. M., & S.C., v. d. S. (2010). *Visualizing pedestrian flows using GPS-tracking to improve inner-city quality*. Paper presented at the Walk21.

Wang, W., Yang, J., & Muntz, R. R. (1997). *STING: A Statistical Information Grid Approach to Spatial Data Mining*. Paper presented at the Proceedings of the 23rd International Conference on Very Large Data Bases.

Xie, K., Deng, K., & Zhou, X. (2009). *From trajectories to activities: a spatio-temporal join approach*. Paper presented at the Proceedings of the 2009 International Workshop on Location Based Social Networks, Seattle, Washington.

Yan, Z., Chakraborty, D., Parent, C., Spaccapietra, S., & Aberer, K. (2011). *SeMiTri: a framework for semantic annotation of heterogeneous trajectories*. Paper presented at the Proceedings of the 14th International Conference on Extending Database Technology, Uppsala, Sweden.

Yan, Z., Parent, C., Spaccapietra, S., & Chakraborty, D. (2010). *A hybrid model and computing platform for spatio-semantic trajectories*. Paper presented at the Proceedings of the 7th international conference on The Semantic Web: research and Applications - Volume Part I, Heraklion, Crete, Greece.

Zhou, C., & Meng, X. (2011). *STS: complex spatio-temporal sequence mining in flickr*. Paper presented at the Proceedings of the 16th international conference on Database systems for advanced applications - Volume Part I, Hong Kong, China. http://dl.acm.org/citation.cfm?id=1997328&CFID=85392474&CFTOKEN=79196120

Zimmermann, M., Kirste, T., & Spiliopoulou, M. (2009). Finding Stops in Error-Prone Trajectories of Moving Objects with Time-Based Clustering. In D. Tavangarian, T. Kirste, D. Timmermann, U. Lucke & D. Versick (Eds.), *Intelligent Interactive Assistance and Mobile Multimedia Computing* (Vol. 53, pp. 275-286): Springer Berlin Heidelberg.