Web Application for Travel Diary Annotation and Methods for Trip Destination and Purpose Inference

ANDERS MAGNUSSON
Abstract

Gathering travel survey data in an automated way reduces user fatigue and non-response rates, which enables longer studies and thus possibly lower survey costs [1]. Access to GPS data from users’ mobile phones and extensive GIS data enables automated inference of trip attributes such as travel mode, trip destination and trip purpose.

This thesis proposes a rule-based method and four machine learning methods for inferring trip destination and two machine learning methods for inferring trip purpose. The trip destination inference methods take the public or personal points of interest, POIs, within a specified radius of the trip end point (the \( R \)-neighborhood) as input and the methods are inverse distance weighted.

**Inverse Distance Weighting** is an IDW-only method method that is static in its learning ability and generic in its generality, meaning that it does not improve as the sample increases and that it does not need user specific data. The method assigns trip destination to the POI in the \( R \)-neighborhood that is nearest to the trip end point.

**Most Frequently Visited Nearest POI** is a dynamic, user specific method that takes a user’s validated previous trip destinations as input and assigns trip destination to the POI in the \( R \)-neighborhood that the user has visited most.

**Inverse Distance Weighted Naïve Bayes Classifier** is a dynamic, general method that takes the day-time period of arrival at destination and the POI types of all users’ previous validated trips as input. The method calculates the probability of each of the POI types in the \( R \)-neighborhood based on the the number of previous visits to each POI type during each day-time period.

**Inverse Distance Weighted Spatial Relevance** is a dynamic, generic method that uses a spatial adaptation of the *text frequency - inverse document frequency* measure that is often used in text mining and information retrieval. The method takes the spatial density of each POI type in the \( R \)-neighborhood and the spatial density of those POI types in the whole data set as input.

**Inverse Distance Weighted Naïve Bayes Classifier with Spatial Relevance** is a dynamic, generic method that combines the spatial relevancy measure and the Naïve Bayes Classifier classifier.

The thesis also proposes a rule-based method and a machine learning method for inferring trip purpose. **Individual Trip History Based Rules** is a user specific and dynamic method that takes a user’s set of validated trips as input and if there are previous trips to the current trip destination, it sets the trip purpose equal to the previous trip purpose of that trip destination.

**Naïve Bayes Classifier**, is a generic and dynamic method that uses the day-time period of the trip, the POI type of the destination and the verified previous trips of all users as input and calculates the probability of each trip purpose based on those parameters.

A web based trip annotation system is designed and implemented to enable the assessment of the methods and to provide an information feedback loop for the trip purpose
inference methods. The trip annotation system is based fully on open source components. A mobile application, which is an early version of the *KTH mobility collector* [9], is used to collect the travel data which is stored in a spatial database.

To assess the accuracy, a small-scale travel study with 9 users is performed and 208 trips are annotated. The accuracy of *Most Frequently Visited Nearest POI* is assessed using the 128 trips where destination is a personal POI and shows an accuracy of 67.2%. The accuracies of the remaining trip destination inference methods are assessed using the 36 trips where destination is a public POI present in the GIS data. The accuracies are 38.9% for *Inverse Distance Weighting*, 36.1% for *Inverse Distance Weighted Spatial Relevance* and 25.0% for *Inverse Distance Weighted Naive Bayes Classifier* and *Inverse Distance Weighted Naive Bayes Classifier with Spatial Relevance*. *Inverse Distance Weighting* outperforms *Inverse Distance Weighted Spatial Relevance*, *Inverse Distance Weighted Naive Bayes Classifier* and *Inverse Distance Weighted Spatial Relevance* but *Inverse Distance Weighted Naive Bayes Classifier* shows improving results as the sample increases.

For the trip purpose inference methods all 208 trips are used to assess the accuracy. *Individual Trip History Based Rules* shows an accuracy of 57.7% and *Naive Bayes Classifier* shows an accuracy of 76.0%. Results show that *Naive Bayes Classifier* is clearly superior to *Individual Trip History Based Rules* and the accuracy of *Naive Bayes Classifier* increases as the sample increases. The increasing accuracies of the dynamic methods indicates the need for a larger-scale study to fully evaluate the accuracy of those methods.
Contents

1 Introduction 4
   1.1 Evolvement of Travel Studies .......................... 4
   1.2 Motivation and Research Direction ....................... 5

2 Related Work 6
   2.1 GPS Based Travel Studies ................................. 6
   2.2 Trip Inference in GPS Based Travel Studies ............. 6

3 Research Methodology 8

4 Trip Annotation System (Components / Design) 10
   4.1 General System Description .................................. 10
   4.2 GIS Data Selection .............................................. 10
   4.3 The Mobility Collector ........................................... 11
   4.4 Trip Validation .................................................. 12
      4.4.1 Deleting Sensitive Trips or Incorrect Trip Points or Removing Fake Trips ........................................... 12
      4.4.2 Merging or Splitting Trips or Trip Legs .................. 12
      4.4.3 Setting Transit POIs and Travel Mode .................... 13
      4.4.4 Verifying or Correcting Inferred Trip Destination or Trip Purpose ........................................... 13
   4.5 Trip Inference .................................................... 14
      4.5.1 Origin Inference ............................................. 14
      4.5.2 Trip Destination Inference .................................. 15
      4.5.3 Trip Purpose Inference ..................................... 20
      4.5.4 Derived Trip Attributes .................................... 20

5 Empirical Evaluations 21
   5.1 Experimental Setup ............................................ 22
   5.2 Trip destination Inference Accuracy Assessment .......... 22
   5.3 Trip Purpose Inference Accuracy Assessment ................ 24

6 Discussions 24

7 Conclusions and Future Work 27
   7.1 Conclusions ....................................................... 28
   7.2 Future Work ....................................................... 28

References 30
# List of Figures

1. Flow chart of the research methodology .............................................. 8  
2. User interface of the trip annotation system ........................................ 11  
3. Removing a fake trip in the user interface .......................................... 13  
4. Setting a transition point in the user interface .................................... 14  
5. Setting the trip destination in the user interface ................................... 15  
6. Illustration of the proposed spatial relevancy measure ........................... 19
## List of Tables

1. Categorization of the inference methods ........................................ 16
2. Day-time periods used in the inference methods ................................. 17
3. Trip destination inference accuracy .................................................. 23
4. Trip purpose inference accuracy ...................................................... 23
5. Confusion matrix for the trip purpose inference method ITH-TPI ........ 25
6. Confusion matrix for the trip purpose inference method NB-TPI .......... 25
1 Introduction

This section first describes how travel studies have evolved from aggregate trip-based models in the 1950s to disaggregate activity based models and then summarizes important research on GPS-based travel studies and inference of trip purpose. Then it gives the motivation for the thesis and presents the research questions.

1.1 Evolvement of Travel Studies

Accurate models of travel behavior are important when analyzing how policies, such as congestion pricing, effect travel behavior. Aggregate trip-based models were first developed in the 1950s and were then evolved to disaggregate trip-based models in the 1980s and 1990s. In recent years these models have gone from being trip-based to being activity-based [5].

Aggregate models, such as the traditional four-step model, uses the trip as the entity of analysis and considers trips between zones. Disaggregate models, also known as second generation travel demand models or discrete choice models, are probabilistic models that also use the trip as the entity of analysis but on the level of the individual traveler. Criticism towards trip-based models has been that the relationship between trips made by an individual traveler is not considered and that trips are often only distributed on a 24-hour period, and sometimes also distributed on periods of the day using time-of-day factors. Trip-based models usually also deal with home-based and non home-based trips in separate models.

Activity-based models uses the activity as the entity of analysis and consider what activities are performed, and where and when they are performed. They deal with the complex interactions between the factors that influence activity choice behavior, the relation between trips, their sequence and their time-of-day.

Traditional data collection in travel surveys has been pen-and-paper surveys where participants are prompted to recollect and account for their trips during a period, often 24 hours. In recent years focus has been on developing more efficient methods for travel surveys, as studies have shown that less need for manual work by respondents reduces the non-response rate and enables longer survey periods [1].

Some studies on the inference of trip attributes such as trip purpose have been performed. Rule based methods together with GIS data, such as a set of POIs, have been used and evaluated and the results have been at similar level as in pen-and-paper studies [12] when a small set of trip purposes has been used. Bohte and Maat [1] evaluated a rule based method by performing a study with 1104 respondents in the Netherlands and achieved an accuracy of between 74% (trip purpose “home”) and 4% (trip purpose “school”). Experiments of using machine learning methods for trip purpose inference has also been made [7].
1.2 Motivation and Research Direction

The shift towards activity-based models has increased the need for accuracy and detail in the data used in travel behavior models, such as what activities individual travelers are performing and when they are performing them, how the activities relate to each other and in what order they are performed. It has also created the need to focus on the decision-making process of travel, something that traditionally has not been the focus of travel studies.

To reduce the amount of manual work for participants in travel studies, automatically inferring trip attributes such as trip purpose has been the focus of several studies. Reducing user fatigue could enable multi-day data collection, which in turn could reduce the cost of travel studies and therefore motivate the development and maintenance costs of a GPS enhanced travel survey system. Inferring trip attributes through minimal user effort is an important step towards automated collection of data for travel studies. Trip purpose and destination are considered to be two key attributes of a trip. Therefore, the research questions this thesis attempts to answer are:

1. Which data and methods are appropriate for trip purpose and destination inference?
2. How can a system based on open-source components that facilitates semi-automatic collection of trips and validated trip attributes be designed?

The methodology of the thesis is that, first, a literature study is performed to give an overview of the current state of research. Then, relevant GIS data that can be used as input for the inference methods is selected. After that, the inference methods are designed and a trip annotation system is designed and implemented to facilitate the testing of the inference methods. After that, a small scale study is performed to provide the basis for the evaluation of the inference methods and lastly the inference methods are evaluated to see how they perform.

This thesis contributes to the current state of research by proposing and evaluating several rule-based and machine learning methods that combine GPS data, GIS data and data on previously performed trips. For the collection of travel data, a mobile application designed for data collection is used. For the validation of travel data and automated inference of trip attributes, a web based trip annotation system based on open-source software is designed and implemented.

The remainder of the thesis is structured so that Section 2 describes the current research on inference of trip purpose and trip destination, Section 3 describes the research methodology, Section 4 describes the developed trip annotation system and the proposed inference methods, Section 5 describes the performed study and the evaluated accuracies, Section 6 discusses the work performed and Section 7 concludes and outlines possible future work.
2 Related Work

This section accounts for interesting research in the field. First, Section 2.1 describes the shift from traditional travel surveys towards GPS based travel studies. Then, Section 2.2 describes the research that has been performed on the inference of trip purpose.

2.1 GPS Based Travel Studies

In the last few years research on travel behavior have moved in the direction of GPS based travel studies. The National Cooperative Highway Research Program’s report Applying GPS Data to Understand Travel Behavior [14] describes the benefits of GPS based travel surveys over traditional travel survey methods as follows:

First, GPS based surveys provide a more accurate and detailed account of the spatial and temporal aspects of personal travel than what survey respondents are able to recall and report, and GPS data sets have been used to correct significant trip underreporting errors associated with pen-and-paper or phone-based activity surveys (Battelle Memorial Institute 1997; Wolf, Bricka, et al. 2004). GPS based surveys should have less respondent burden for capturing travel details by leveraging passive GPS data collection while collecting more information and more accurate information. In addition, by further reducing respondent burden through the use of automated trip purpose, location, timing, and travel mode identification routines, GPS based prompted-recall surveys allow for more complex questions to be asked.

Commonly performed types of GPS based surveys performed in recent years are surveys where trip attributes such as trip purpose are entered manually by the user through a user interface [2, 13] and surveys where trip attributes are inferred automatically [1, 7, 10, 12]. The latest generation of GPS based surveys include GPS-only studies where basic household information has been collected first [14].

Previous surveys that have been studied in this research include the study by Frignani et al. [5] study where GPS data was collected from users during 14 consecutive days. At the end of each day users visited the project website to upload their GPS data and answer travel surveys, which included questions about their activity planning and scheduling processes. The results of the study indicated satisfactory response rates, higher data quality than in other types of travel studies and inclusion of trips often missed in other studies. Feedback on user fatigue also indicated that this type of method can be useful for longer study periods.

2.2 Trip Inference in GPS Based Travel Studies

The inference of trip destination has in previous studies been regarded mostly as an intermediate step in the inference process of trip purpose. Previous studies that have inferred trip destination by matching GIS data with trips have used primitive methods such as
simply matching the trip trajectory’s end point to the nearest POI or parcel with land use data \cite{1,7,12}, likely because the sets of POIs used in those studies have been limited, so more sophisticated methods for trip destination inference have probably not been necessary to use.

Trip purpose cannot be directly deduced from GPS trajectories. Instead, it has to be inferred from GPS trajectories combined with other information, such as GIS data or socio-demographic data. Inference through combining these types of data has commonly been done using rule based methods. Stopher et al. \cite{12} used a method based on GPS logs, GIS data on land use and geocoded locations of home, work / school and the two most frequently visited grocery stores / supermarkets that were manually entered to classify trip purposes. They found that:

If trip purposes are to be defined only to a small set, such as home-based work, home-based education, home-based other, non-home-based (work), and non-home-based (nonwork), then the address information we collect, together with an examination of the frequency over each week that the places are visited and the length of time spent there will largely identify all the trip purposes. It is only if the detail to be obtained on trip purpose is greater than this that significant effort is required to identify purposes.

For more disaggregated trip purposes they also used land use data. They concluded that land use data was essential for the process and that they were able to achieve “a similar level of accuracy to that achieved in self-report diaries” using the method.

Bohte and Maat \cite{1} used a method based on GPS logs, GIS data consisting of a limited set of POIs, and individual user characteristics to infer trip purposes. A set of eight trip purposes was used: “work”, “study”, “shop”, “social visit”, “recreation”, “home”, “other”, and “all purposes”. They allocated the trip end point to the nearest POI within 50 meters, or to home or work points if either of those were within 100 meters. Trips where there were no POIs or home or work points was classified as “unknown”, forcing the user to set trip purpose manually during a validation process. Their method also had a user specific component that classified trips ending within 50 meters of a reclassified “unknown” trip with the reclassified trip’s purpose. To validate they performed a large-scale study where 1104 respondents were selected from three municipalities in the centre of the Netherlands to participate in a one-week study. Bohte and Maat were able to correctly classify the trip purpose “home” at an accuracy of 74%. For the remaining trip purposes, the classification accuracy was between 43% (“all purposes”) and 4% (“school”), which leaves room for improvement. Like \cite{1,12} the herein proposed methods use POIs but extends by using a large set. The proposed methods consist of a rule based method similar to the one in \cite{1} but also more sophisticated machine learning methods.

A few studies where machine learning methods have been used to infer trip purpose have been performed in recent years. Li, Dai, Sahu and Naphade \cite{7} demonstrated a semi-automatic trip analysis system that used a rule based method for initialization and an trip purpose classifier which used user corrected trip purposes as training data. It was

\footnote{a point of interest, POI, is a specific location that some / many may find interesting or useful}
Figure 1: Flow chart of the research methodology

not disclosed in the article what type of classifier that was used, but both the rule based method and the trip purpose classifier used arrival time, arrival day, stop time, nearest POI, and distances between trip origin / trip destination and home / work / school as inputs. They concluded that further evaluation over a larger population would be needed to assess the accuracy of the system. Like [7] the herein proposed methods are machine learning methods and are, in addition to [7], evaluated through a small scale study.

3 Research Methodology

This section describes how the research is structured and performed and Figure 1 shows a flow chart describing the process. The trip annotation system is designed to be part of a semi-automatic travel diary collection system that leverages current mobile and web-GIS capabilities to allow the seamless collection (using a mobile collector) and the periodic, convenient travel-diary-annotation of trips of its users. The purpose of the trip annotation system is to allow users to visualize, validate / correct and annotate their trips through an intuitive User Interface, UI, while trip destination and purpose inference is performed by the system in the background. Feedback between the validation and the inference allows
the system to use validated trip data when inferring trip destination and purpose. The mobile collector that is used together with the trip annotation system is an early prototype of the Mobility Collector [9], a mobile application that the trip annotation system is integrated with through a shared spatial database. The design and implementation process of the trip annotation system and the inference methods is performed in the following steps, which are illustrated in Figure 1:

**STEP 1: A literature study on travel surveys is performed.** The focus of the literature study is on research on the inference of trip attributes. The trip attributes that are considered interesting and relevant to study are selected.

**STEP 2: Relevant GIS data is selected.** Detailed GIS data is considered to be a crucial input for methods that infer trip destination and purpose. In this study an extensive source of POI data is used. The data should be free to use, complete, up-to-date and possible to download or to perform spatial queries on through an API. The POIs should also be categorized based on their type (e.g. restaurant, shop, etc).

**STEP 3: The inference methods are designed.** This step is a function of the literature study and the available GIS data to use as inputs for the inference methods. The aim of the inference task is to infer trip destination and purpose, as they are considered to be two key attributes of a trip and cannot be inferred directly from the GPS trajectory. The literature study shows that promising results in inferring trip purpose have been obtained using both machine learning and rule based methods. Therefore, both of these types of methods are designed and implemented.

**STEP 4: A trip annotation system is designed and implemented.** A system functionality requirement analysis is performed. The database and database tables are created and the POI table is populated. Mockups of the UI are created, which are then used as support in the implementation. Background maps from Google are implemented as WMS\(^2\). The inference methods are implemented as database functions that performs analysis on validated trip data from the database.

**STEP 5: A small scale study is designed and executed.** During the study, the users install and keep the Mobility Collector on their Android devices. They are instructed to log on to the trip annotation system every evening to validate their trips regarding stop / transition periods and trip destination and purpose, and to set the travel mode\(^3\) for each trip leg.

**STEP 6: The inference methods are evaluated.** The inferred trip destinations and purposes of each inference method are compared with the validated trip destinations and purposes, and relevant metrics are calculated. The trip purpose inference methods are evaluated using all validated trips that are associated with a POI, whereas the trip destination inference methods are evaluated using a subdivision of scenarios based on the

---

\(^2\)The Open Geospatial Consortiums web service standard for georeferenced maps in raster format

\(^3\)Travel mode was not inferred by the early prototype of the Mobility Collector used in the study
methods applications. Different subsets of the validated trips are used for the different scenarios.

4 Trip Annotation System (Components / Design)

The following sections describe the Mobility Collector, the GIS data used in this study, the functionality for validating trips and inferred trip attributes by users and the inference methods that are designed and implemented.

4.1 General System Description

The collection system consists of two parts: a mobile collector and an trip annotation system. The mobile collector is a mobile application that collects GPS data, segments the data into trips and uploads it to a database that it shares with the trip annotation system. The mobile collector is described more later in this section. The trip annotation system is a web application for annotation and inference that implements the inference methods that are proposed in this thesis. It also implements functionality for verifying, editing and adding attributes to trips. Feedback is built into the system, which allows the inference methods to use user-verified and annotated trip data in the inference process.

The trip annotation system is implemented as a web application using open-source software. The web interface is implemented using W3C’s\textsuperscript{4} standard technologies for serving and rendering web pages (HTML, CSS and JavaScript) and OpenLayers is used to display maps and geographic data. Microsoft’s ASP.NET and Model-View-Controller frameworks are used, with C# as the server-side programming language. The application runs in a Linux / Apache environment with Mono\textsuperscript{5}, which enables a Microsoft .NET application to run on a Linux server. Data about trips and users is stored in a PostgreSQL object-relational database, spatially enabled with PostGIS.

4.2 GIS Data Selection

A review of the freely available sources of POIs is performed so that a suitable data source can be be selected, which shows that the OpenStreetMap (OSM) data, which contains an extensive set of POIs, is the most complete and up-to-date dataset available. Therefore the OSM data is selected as the primary source of geographic data.

The OSM data is classified by the contributors using tags that consist of key-value pairs, such as “amenity=restaurant”, “amenity=university”, “shop=supermarket”, etc. When the dataset is used in this study, a subset of the tags considered relevant for the study is selected and the values of those keys are used as POI types. A few examples of the POI types are restaurant, fast-food, pub, university and supermarket. The selection from the full set of OSM POIs is done so that the list

\textsuperscript{4}the World Web Consortium, http://www.w3.org

\textsuperscript{5}http://www.mono-project.com
for users to select from when manually setting the POI type of the trip destination should be reasonably small. The selection is done through a manual review of the OSM POI types, where types that seem uncommon or irrelevant are left out. A set of personal POI types is also defined and used. Those POI types are home, work, friend / family and other. As the study were to be set in Sweden, only the POIs in Sweden are selected and used. After the selection process, the selected subset of the OSM POIs are imported into the PostgreSQL / PostGIS database.

4.3 The Mobility Collector

The mobile collector used in this research and the reported trial is based on an early prototype of the Mobility Collector [9] that in a power-efficient manner samples the GPS position of the mobile terminal / smart phone of the user at equal distance intervals and stores the equidistance GPS trajectory of the user in the database. Additionally, based on motion characteristics derived from location and accelerometer measurements the collector detects potential transition periods and stop periods which respectively correspond with locations in the GPS trajectory at which the user has possibly changed transport mode and finished a trip (i.e., trip destination point).

The detection of transition periods, which is where the user has changed travel mode, and stop periods, which is where the user has ended a trip, are based on motion characteristics that are derived from location and accelerometer measurements. A time-based differentiation between transition (<2min) and stop periods (>2min) is used.
4.4 Trip Validation

The users validates trips and inferred trip attributes in the trip annotation system. This is performed so that feedback is provided to the inference methods and so that data is collected for validation of the inference methods and for further research. The users also manually set trip attributes that are not automatically inferred. More specifically, the major functionality of the UI (see Figure 2) allows for:

- Deleting sensitive trips or incorrect trip points or removing fake trips
- Merging or splitting trips or trip legs
- Setting transit POIs and travel mode
- Verifying or correcting inferred trip destination or trip purpose

The following sections describe the validation functionality in detail.

4.4.1 Deleting Sensitive Trips or Incorrect Trip Points or Removing Fake Trips

To protect user privacy, trips that the user perceives as sensitive can be deleted by the user. Similarly, fake trips, which are incorrectly registered trips due to noise from the collector, can also be deleted through a user action (see Figure 3). Trips deleted due to privacy concerns are deleted from the database while removed fake trips are hidden from the user, but remains in the database for analysis purposes.

A sequence of consecutive trip points between two consecutive stop periods is referred to as a trip (trajectory). A sequence of consecutive stop or transition periods is referred to as a trip leg (trajectory). Trip points are the GPS records or average GPS records visualized as nodes in the trip trajectory. Incorrect trip points, due to inaccurate position determination by the collector, are easily identified by the user through visual inspection of the trip and can consequently be deleted through a user action.

4.4.2 Merging or Splitting Trips or Trip Legs

Incorrect trip segmentation by the collector can lead to the following inaccuracies:

- A single trip registered as multiple trips
- Multiple trips registered as a single trip
- Incorrect trip length

The first inaccuracy can occur when movement is slow, for example when the user is stuck in a traffic queue. That scenario can lead to failure of the collector to register the stop point and instead merging the trip with the following trip. The second inaccuracy can occur when the stop time at a trip destination is short which can make the collector misinterpret the stop period for a transition period.

To enable correction of these inaccuracies, the UI allows the user to merge and split trips by changing the period between three possible types (regular, transition, stop) that segment the GPS trajectory into trips and trip legs.
4.4.3 Setting Transit POIs and Travel Mode

Travel mode is manually set by the user for each trip leg, through selection from a fixed list of nine travel modes: walk, bicycle, car, bus, subway, train, ferry, tram and other. These travel modes are included as they are considered to be the most common modes in the area of the study.

The user is also able to connect a transport related POI to each transition point, through selection from a list. For each transition point, the list is generated from a query that returns the transport related POIs (e.g., bus stops, train stations, subway stations) within a certain radius of the $R$-neighborhood of the transit point (see Figure 4). This radius is denoted $R_{\text{trans}}$.

4.4.4 Verifying or Correcting Inferred Trip Destination or Trip Purpose

Trip destination and purpose are automatically inferred by the system. These trip attributes must be verified by the user and can be corrected through selection from prioritized lists (see Figure 5) and maps. The user can also create a new POI, which is necessary if the trip destination is missing from the set of available POIs.
4.5 Trip Inference

Correctly inferring trip origin and destination is important because they are interesting attributes of a trip, but also because they can be used in inference of other key attributes, such as trip purpose which is the focus of many travel surveys. Trip destination is used as input for the trip purpose inference methods proposed in this study. The system therefore infers trip origin, trip destination and trip purpose of each collected trip, using trip trajectories, POIs from the OSM data and verified previous trips. The following sections describe in detail the methods proposed in this study.

4.5.1 Origin Inference

Origin inference is considered to be a trivial task, as a user’s current trip origin always corresponds to the user’s previous trip destination, unless, for example it is the user’s first trip. Other examples of where the user’s trip origin will not correspond to the user’s trip destination is when a trip is following a missing trip or a trip that has been deleted by the user due to privacy concerns. A simple rule based method is therefore used to infer trip
origin. Origin is not validated by the user, therefore the accuracy of this method is not evaluated.

The origin inference method sets origin to the previous trip destination if (1) the previous trip has a trip destination POI and (2) that POI is within a certain radius of the current trip’s start point, denoted $R_{\text{origin}}$. Else, the current trip origin is annotated as unknown.

4.5.2 Trip Destination Inference

Human behavior is considered to be predictable and regular. Previous studies have shown high regularity in human movement \cite{11} and the methods used for inferring trip destination are based on this assumption. The regularity can be within the behavior of the individual or within a group of individuals. To assess which regularity is a better predictor for the travel behavior (trip destination, trip purpose) both methods that are based on the aggregate travel behavior as well as methods that are based on the individual’s travel behavior are devised and assessed.
<table>
<thead>
<tr>
<th>Learning ability</th>
<th>User specific</th>
<th>Generality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Static</td>
<td>User specific</td>
<td>TDI, SR-TDI</td>
</tr>
<tr>
<td>Dynamic</td>
<td>VF-TDI, ITH-TPI</td>
<td>NB-TDI, SRNB-TDI, NB-TPI</td>
</tr>
</tbody>
</table>

Most trips’ origins and trip destinations are considered to be either personal (home, work, friend / family or other) or public POIs. Use of an extensive POI source, such as the OSM data, increases the possibility of accurately inferring trip origin and trip destination, but also leads to potential difficulties. For example, high concentration of POIs, which often occur in dense urban areas, can lead to having many POIs clustered together within a trip end point’s $R$-neighborhood, which is defined as the area within a certain radius, $R_{dest}$, of the trip end point. This, together with the accuracy of the collector’s GPS, increases the need for more intelligent trip destination matching methods. The postulated pros and cons of using an extensive set of POIs is discussed in Section 6.

The following sections propose five trip destination inference (TDI) methods. All TDI methods are inverse distance weighted, so IDW is omitted from the abbreviations for readability. The methods require that there are public or personal POIs in the trip end point’s $R$-neighborhood. If there are no POIs in the $R$-neighborhood, the user can create a new POI or set trip destination as not a POI.

The proposed methods are categorized (see Table 1) according to their learning ability (static over time vs dynamically improving over time as the training sample / evidence increases) and their generality (user specific vs generic method). The static methods are considered more suitable for shorter studies as they can give results without having to build up a classifier, while dynamically improving models are considered more suitable for longer studies where the classifier can build up and produce better classifications over time. These postulated characteristics are evaluated in Section 5.

**VF-TDI: Most Frequently Visited Nearest POI** Visited POIs are either personal POIs (home, friend / family, work or other) or public POIs (e.g. restaurant, school, supermarket, etc) that the concerned user has visited. If there are visited POIs in the trip end point’s $R$-neighborhood, then the POI that has the most previous visits by the user is set as trip destination. Smallest distance between trip end point and POI is used for tie-breaking.

**TDI: Inverse Distance Weighting** To account for distance between the trip end point and the POIs in the trip end point’s $R$-neighborhood, a distance function $w(d(e,p))$ is used. The function $w(d(e,p))$ of $e$ depends on distance $d(e,p)$ between POI $p$ and trip end point $e$, and is calculated as:

$$w(d) = \begin{cases} 1 & \text{if } d(e,p) < 1 \text{ meter} \\ \frac{1}{d(e,p)^2} & \text{if } 1 \leq d(e,p) \leq R_{dest} \end{cases}$$

(1)
Table 2: The day-time periods used in the Naïve Bayes based inference methods (NB-TDI, SRNB-TDI and NB-TPI)

<table>
<thead>
<tr>
<th>ID</th>
<th>Time Interval</th>
<th>Day Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>04:00 - 12:00 &quot;Morning&quot;</td>
<td>Weekday</td>
</tr>
<tr>
<td>2</td>
<td>12:00 - 20:00 &quot;Afternoon&quot;</td>
<td>Weekday</td>
</tr>
<tr>
<td>3</td>
<td>20:00 - 04:00 &quot;Evening&quot;</td>
<td>Weekday</td>
</tr>
<tr>
<td>4</td>
<td>04:00 - 12:00 &quot;Morning&quot;</td>
<td>Weekend day</td>
</tr>
<tr>
<td>5</td>
<td>12:00 - 20:00 &quot;Afternoon&quot;</td>
<td>Weekend day</td>
</tr>
<tr>
<td>6</td>
<td>20:00 - 04:00 &quot;Evening&quot;</td>
<td>Weekend day</td>
</tr>
</tbody>
</table>

where $R_{dest}$ is the radius of the trip end point $e$’s $R$-neighborhood. The trip destination is set to the POI in the $R$-neighborhood that maximizes $w(d)$ of $e$.

**NB-TDI: Inverse Distance Weighted Naïve Bayes Classifier**  This classifier is based on the Bayes rule: $P(B|A) = P(A|B) \times P(B)/P(A)$. This allows for calculation of the probability of the class $B$ given the observed variable $A$ based on the previously observed prior probabilities of the attribute and the class as well as the class-conditional probabilities of the attribute. The Naïve Bayes classifier assumes that the observed variables are conditionally independent $P(B|A_1...A_n) = P(B|A_1) \times ... \times P(B|A_n)$ which allows to calculate / estimate the class probabilities based on each variable separately and then simply multiply. The strength of the Naïve Bayes classifier is that it needs to estimate far fewer parameters from the same training data then a full Baysian classifier and because in real word classification scenarios the dependence between the variables is indeed to a large degree encoded in the class variable. Naïve Bayes classifiers have shown to achieve good real-world results even when the assumption of conditionally independent variables is not true.

NB-TDI uses the day-time period of arrival at the trip destination (see Table 2) as the predictor variable. The probability of each destination type that is present in the $R$-neighborhood of the trip end point is calculated from the prior probabilities of the destination type, the day-time period, and the probability of the day-time period given the destination type based on previous observations. This probability is then combined with the IDW and the trip destination is set to the POI in the $R$-neighborhood that maximizes $P_{NB}(dt)$. All verified trips of all users are used to calculate the probabilities, which makes the classifier general and dynamically learning.

**SR-TDI: Inverse Distance Weighted Spatial Relevance**  The distribution of POIs over space is non-uniform. Areas such as central parts of cities often have higher POI density than suburban parts and there are often neighborhoods within central parts that have higher POI density than other more residential areas. It is also common that POIs of the same type are located near each other, forming so called clusters.

A hypothesis is that if the spatial density of a POI type represented in the $R$-neighborhood is higher than expected, it is more likely that the trip is linked to a POI of that type.
instead of to a POI of a type that has expected spatial density. A spatial relevance factor is introduced to account for this.

The spatial relevance factor is an adaptation of text frequency - inverse document frequency, **tf-idf**, relevancy measure to a spatial setting. **tf-idf** is an often used weighting factor in text mining and information retrieval and shows the importance of a term to a document, in a collection of documents [16]. The first part in **tf-idf**, term frequency, denoted $tf(t, d)$ for document $d$ and term $t$, is a measure of the relative frequency of a term in a document. It is defined as:

$$tf(t, d) = \frac{f(t, d)}{\max\{f(w, d) : w \in d\}}$$  \hspace{1cm} (2)

The second part, inverse document frequency, denoted $idf(t, D)$, is a measure of how frequent a term is across all documents. The purpose of $idf(t, D)$ is to reduce the effect of terms that occur too often in documents in the collection to be useful in relevancy determination. It is defined as:

$$idf(t, D) = \log\frac{|D|}{|\{d \in D : t \in d\}|}$$  \hspace{1cm} (3)

where $t$ is the term, $|D|$ is the total number of documents in the collection and $|\{d \in D : t \in d\}|$ is the number of documents containing the term $t$. The measure **tf-idf** is a combination of the two parts and is defined as:

$$tf - idf(t, d, D) = tf(t, d) \times idf(t, D)$$  \hspace{1cm} (4)

When adapting **tf-idf** to a spatial setting, document $d$ corresponds to the trip end point’s $R$-neighborhood, the collection of documents $D$ correspond to a uniform rectangular grid of nine cells. The terms corresponds to the POI types.

As a simplified example of the spatial adaptation of **tf-idf**, consider a study area divided into a uniform rectangular grid of nine cells. There are two POI types, *restaurant* and *grocery store*, which both have uniform prior probabilities of being the trip destination, which means that the average user visits a *restaurant* and a *grocery store* equally often. The grid and the spatial distribution of the POIs are illustrated in Figure 6. All trips are assumed to have a POI as the trip destination and instead of using the trip end point’s $R$-neighborhood, the cell where the trip ends is used. Consider the following cases:
Figure 6: Illustration of the proposed spatial relevance measure based on tf-idf for two POI types (R and G). While the relative frequencies of types in the upper right cell and the lower left cell are exactly opposite, because of the spatial distribution of the types in the study area (all the grid cells) the spatial relevance of the two types are similar to each other in the upper right cell while the spatial relevance of G is much higher than the spatial relevance of R in the lower left cell.

Case one: A trip ends in the top-right cell where there are four restaurants and one grocery store. While restaurant is more frequent in the cell where the trip ends and therefore has higher \(tf\) than grocery store, there are also more cells were a restaurant can be visited. This yields lower \(idf\) for restaurant than for grocery store, giving restaurant and grocery store similar tf-idf.

Case two: A trip ends in the bottom-left cell where there are one restaurant and four grocery stores. More grocery stores than restaurants in the cell where the trip ends and lower number of cells where grocery store is represented yields both higher \(tf\) and \(idf\) for grocery store than for restaurant. This results in much higher tf-idf for grocery store, making it more probable that the trip destination is one of the grocery stores than the restaurant.

Based on the frequency distributions in the top-right and bottom-left cells, one would expect symmetric relevance values. The asymmetry in the spatial relevance arises because the measure considers the distributions of the types in the entire study area.

The spatial tf-idf measure is calculated for each POI type represented in the trip end point’s \(R\)-neighborhood. The measure is scaled to range between zero and one by dividing tf-idf for each POI type in a cell with the sum of tf-idf of all POI types in the cell, so that it has approximately the same weight as the IDW. The scaled tf-idf measure is then combined with the IDW, and the trip destination is set to the POI in the \(R\)-neighborhood that maximizes \(tf - idf \cdot w(d(e, p))\).
SRNB-TDI: Inverse Distance Weighted Naïve Bayes Classifier with Spatial Relevance To account for the probability of visit to each POI type based on all previous trips, as well as the spatial distribution of the POIs, the Naïve Bayes Classifier and the spatial adaptation of \( \text{tf-idf} \) and the IDW are combined. Trip destination is set to the POI in the trip end point’s \( R \)-neighborhood that maximizes \( tf-idf_{[0,1]} \times P_{NB}(dt) \times w(d(e,p)) \).

4.5.3 Trip Purpose Inference

Two trip purpose inference methods are proposed, ITH-TPI which is user specific and dynamic and NB-TPI which is generic and dynamic. Trip purpose inference is performed after the user has validated trip destination for a trip, allowing ITH-TPI and NB-TPI to use trip destination as input. Six trip purposes are defined and used: Worktrip, Pick up / drop of, Leisure, Shopping, Other and Home. The following sections describe the TPI methods.

ITH-TPI: Individual Trip History Based Rules If the user’s set of previous trips contains a corresponding trip, which is defined as a user-validated trip that has the same trip destination as the current trip, then the trip purpose is set equal to the trip purpose of the corresponding trip, as it is considered likely that a user’s trips that have the same trip destination also have the same trip purpose, regardless of the trips origin, time of departure, time of arrival or other attributes of the trips. If no corresponding trip is found, the method does not infer. ITH-TPI is applied after the trip destination has been validated by the user and ITH-TPI assumes that the trip destination is a personal or public POI (does not infer otherwise).

NB-TPI: Naïve Bayes Classifier A Naïve Bayes classifier is used to infer trip purpose. The classifier uses all validated trips from all users as training data to build the classifier, allowing the method to use validated trip attributes. The following predictor variables are used:

- Day-time period of arrival at the trip destination (see Table 2)
- Validated POI type of the trip destination

Similar to ITH-TPI, NB-TPI is also applied after the trip destination has been validated by the user and assumes that the trip destination is a personal or public POI (does not infer otherwise).

4.5.4 Derived Trip Attributes

Some attributes are derived directly from the trip trajectory without the need of additional data. These attributes are:

- Time of departure
- Time of arrival
• Node time
• Link speed

These attributes are made visible to the user as attribute information on trip points or links in the UI. The main reason for this is to highlight potential problem areas for the user, e.g., a too short node time at a destination which might indicate some trip segmentation problem or a too large or small link speed which might indicate a GPS error that results in an unrealistic computed speed.

5 Empirical Evaluations

A study was performed to assess the accuracy of the TDI and TPI methods. The accuracy for the rule based method VF-TDI is assessed using the trips where trip destination is annotated with a personal POI, whereas TDI, NB-TDI, SR-TDI and SRNB-TDI are assessed using the trips where trip destination is annotated as a public POI from the OSM data. The reason for evaluating the methods based on different subsets of the data is that VF-TDI is the only method that has a chance of inferring personal POIs, while the method’s chance of inferring public POIs in the short run is limited, as the POIs that the method has access to are those that the user has visited. In the longer run, when the user revisits public POIs, the method should also be able to predict for those trips, but due to the often short survey periods used in travel surveys such long term benefits of VF-TDI have little utility. TDI, NB-TDI, SR-TDI and SRNB-TDI all require public POIs and when they have no data to perform predictions on, then the analysis of their predictable NULL result is not interesting. The accuracies of the TDI methods are evaluated later in this section.

Although accurate “geocoding” of trip destination is important, previous studies did not focus on this, hence there are no evaluations of the inference of trip destination in previous studies for comparison purposes.

The accuracies for the TPI methods are evaluated excluding the trips that were annotated with not a POI, as the methods are designed to function when the trip destination is either a public or a personal POI, but not when the trip destination is annotated with not a POI. The accuracies of the TPI methods are evaluated later in this section.

The study took place between 2013-11-12 and 2013-11-24 and 9 users participated. In total, 208 trips were recorded with the collector and annotated with the web application. According to the annotations that the users made, out of the 208 trips there were:

• 128 trips where the trip destination was a personal POI
• 39 trips where the trip destination was a public POI not present in the OSM data or outside the $R$-neighborhood

---

6Link speed is the speed between two consecutive nodes.
7The term “geocoding” in the transport science / travel survey domain specifically refers to the recording of the coordinates of a location that is often but not always associated with a place name or address.
8Not all 9 participants were active during the entire period.
• 36 trips where the trip destination was a public POI present in the OSM data
• 5 trips where the trip destination was annotated with not a POI

5.1 Experimental Setup

In the study, the inference methods are configured so that $R_{\text{origin}}$, the radius around the trip origin, is set to 150 m. $R_{\text{dest}}$ and $R_{\text{trans}}$, the radii around the trip destination point and the possibly occurring transition points are both set to 100 m. The $idf$ cell size that is used in SR-TDI is set to 10 x 10 m.

5.2 Trip destination Inference Accuracy Assessment

The accuracy is evaluated for each of the five TDI methods and this section reviews the accuracy of each method one by one.

VF-TDI inferred trip destination correctly for 67.2% of the 128 trips where trip destination was annotated as a personal POI (see Table 3 for TDI accuracies). Grouping the trips by the type of POI that the trip destination was annotated with show that:

• 71 trips were annotated as to home
• 51 were annotated as to work
• 5 trips were annotated as to friend / family
• 1 trip was annotated as to other

In order for VF-TDI to be able to infer the trip destination correctly the user must have previously visited the POI that is the trip’s destination. Therefore, it is reasonable to analyze how VF-TDI performs as each user’s first trip to each trip destination is excluded. When the analysis of this subset of trips, the revisits, was made, the inference accuracy increased to 86.9%. The accuracy of VF-TDI for the trips that were annotated with “home” was 90.3% and the inference accuracy of the trips that were annotated with work was 65.1%. This shows that VF-TDI is a dynamically learning method that improves as the sample for each user increases. Assessing the accuracy of VF-TDI using both the trips that were revisits to personal POIs and those that were revisits to public POIs yielded similar accuracy, 86.8%.

The results for TDI show that it inferred accurately for 38.9% and incorrectly for 61.1% of the 36 trips where trip destination was annotated with a public POI that was present in the OSM data.\(^9\) This indicates that it is not sufficient to simply set the trip destination to the nearest POI of the last GPS record of a trip. The approximately equidistance (50m) location sampling of the collector with occasional GPS fallouts and cold starts during dynamic movement in urban canyons and noisy or missing GPS signal in indoor environments, all collectively resulted in a relatively inaccurate trip end point measurements. This was confirmed when calculating the distance between each trip end point’s position and position of the corresponding POI. Two trips had distances of 14358

\(^9\)TDI and SR-TDI does not give noRes result when there are public POIs in the $R$-neighborhood.
Table 3: Trip destination inference accuracy. # are the number of trips in the subset used to assess the accuracy, TP (true positive) are the correctly inferred trip destinations, FP (false positive) are the incorrectly inferred trip destinations and noRes (no result) are the number of trips where the methods did not infer.

<table>
<thead>
<tr>
<th>Method</th>
<th>#</th>
<th>TP</th>
<th>FP</th>
<th>noRes</th>
<th>% TP</th>
<th>% FP</th>
<th>% noRes</th>
</tr>
</thead>
<tbody>
<tr>
<td>VF-TDI</td>
<td>128</td>
<td>86</td>
<td>13</td>
<td>29</td>
<td>67.2</td>
<td>10.2</td>
<td>22.7</td>
</tr>
<tr>
<td>TDI</td>
<td>36</td>
<td>14</td>
<td>22</td>
<td>0</td>
<td>38.9</td>
<td>61.1</td>
<td>0.0</td>
</tr>
<tr>
<td>NB-TDI</td>
<td>36</td>
<td>9</td>
<td>20</td>
<td>7</td>
<td>25.0</td>
<td>55.6</td>
<td>19.4</td>
</tr>
<tr>
<td>SR-TDI</td>
<td>36</td>
<td>13</td>
<td>23</td>
<td>0</td>
<td>36.1</td>
<td>63.9</td>
<td>0.0</td>
</tr>
<tr>
<td>SRNB-TDI</td>
<td>36</td>
<td>9</td>
<td>20</td>
<td>7</td>
<td>25.0</td>
<td>55.6</td>
<td>19.4</td>
</tr>
</tbody>
</table>

Table 4: Trip purpose inference accuracy

<table>
<thead>
<tr>
<th>Method</th>
<th>#</th>
<th>TP</th>
<th>FP</th>
<th>noRes</th>
<th>% TP</th>
<th>% FP</th>
<th>% noRes</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITH-TPI</td>
<td>208</td>
<td>120</td>
<td>2</td>
<td>86</td>
<td>57.7</td>
<td>1.0</td>
<td>41.3</td>
</tr>
<tr>
<td>NB-TPI</td>
<td>208</td>
<td>158</td>
<td>12</td>
<td>38</td>
<td>76.0</td>
<td>5.8</td>
<td>18.3</td>
</tr>
</tbody>
</table>

and 1290 meters, respectively, and were regarded as outliers and therefore excluded from the analysis. For the remaining 201 trips the distances were between 220.2 and 0.3 meters, with a mean value of 52.4 meters and a standard deviation of 36.6 meters. This additional distance can greatly influence the distance based TDI methods. The completeness and positional accuracy of the OSM POI data might also have influenced the results of the TDI methods. This is discussed in Section 6.

The results for NB-TDI show that it inferred trip destination correctly for 25.0% of the 36 trips where trip destination was annotated with a public POI present in the OSM data. As the sample of 36 trips is small and NB-TDI is a dynamically learning method, the results are likely to improve as sample size increases [4]. Assessing the accuracy of the first and the second half of the 36 trips separately strengthens this theory, as the accuracy was 16.7% for the first half and 33.3% for the second half.

The results for SR-TDI show that the combined IDW and spatial relevancy method inferred trip destination correctly for 36.1% and incorrectly for 63.9% of the 36 trips where trip destination was annotated with a public POI present in the OSM data. This was slightly lower, one less correctly inferred trip destination, than the accuracy of TDI, which is IDW only. Analyzing the prediction results shows that the spatial relevancy component were able to infer the correct POI type 38.9% of the time.

SRNB-TDI inferred trip destination correctly for 25.0% of the 36 trips where trip destination was annotated with a public POI present in the OSM data. SRNB-TDI is a combination of IDW, the Naïve Bayes Classifier from NB-TDI and the spatial relevancy from SR-TDI, and the results show that it is highly influenced by Naïve Bayes Classifier from NB-TDI, as the prediction choices of SRNB-TDI and NB-TDI are identical.
5.3 Trip Purpose Inference Accuracy Assessment

Table 4 shows that ITH-TPI inferred trip purpose correctly for 57.7%, incorrectly for 1.0% of the trips and did not infer for 41.3% of the 208 trips. When the 5 trips where destination was annotated as not a POI are excluded from the analysis, as the design of ITH-TPI and NB-TPI requires that the trip destination is a POI, ITH-TPI inferred trip purpose correctly for 59.1%, incorrectly for 0.5% of the trips and did not infer for 40.4% of the 203 trips. Assessing the first and the last half of the trips, ordered by timestamp, separately showed that the accuracy improved significantly for the last half of the trips. For the first 102 trips, ordered by timestamp, ITH-TPI inferred correctly for 47.1%, incorrectly for 0.0% and did not infer for 52.9% of the trips, while for the last 101 trips ITH-TPI inferred correctly for 71.3%, incorrectly for 1.0% and did not infer for 27.7% of the trips. For larger and longer samples the improvements over times are expected to be even greater but will likely saturate after a certain accuracy [11]. This improvement is likely statistically significant but this assessment is beyond the scope of this thesis.

NB-TPI inferred trip purpose correctly for 76.0%, incorrectly for 5.8% and did not infer for 18.3% of the 208 trips. When the 5 trips where destination was annotated as not a POI are excluded from the analysis, NB-TPI inferred trip purpose correctly for 77.8%, incorrectly for 5.4% of the trips and did not infer for 16.7% of the 203 trips. Assessing the first and the last half of the trips separately showed that, as for ITH-TPI, the accuracy significantly improved for the last half. For the first 102 trips, NB-TPI inferred correctly for 69.6%, incorrectly for 3.9% of the trips, while for the last 101 trips NB-TPI inferred correctly for 86.1%, incorrectly for 6.9% of the trips. This is also likely significant but the analysis is outside the scope of this thesis.

Trip purposes home and worktrip were inferred with high accuracy by both ITH-TPI and NB-TPI. ITH-TPI inferred home and worktrip correctly for 81.9% respective 71.9% of the trips, whereas NB-TPI performed slightly better at 90.3% respective 87.7% of the trips. Trip purposes leisure, shopping and other were inferred with lower accuracy by both ITH-TPI and NB-TPI. ITH-TPI inferred leisure, shopping and other correctly for 38.2%, 25.0% respective 9.1% of the trips, whereas NB-TPI performed better at 67.7%, 62.5% respective 45.5% of the trips. Both methods failed to infer trip purpose pick up / drop off, likely because there were too few trips with that observed trip purpose. See Table 5 and Table 6 for confusion matrices of the inferred trip purposes of ITH-TPI and NB-TPI.

The results show that NB-TPI is superior to ITH-TPI. They also show that both methods are dynamically learning methods where the results improve over time, as they show better accuracy for the second half of the trips compared to the first half.

6 Discussions

VF-TDI shows promising results in inferring POIs that have been previously visited, as the accuracy for revisits to both personal and public POIs were 86.8%. It needs, however,
Table 5: Confusion matrix for the trip purpose inference method ITH-TPI

<table>
<thead>
<tr>
<th>Observed</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>noRes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.Worktrip</td>
<td>41</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>16</td>
<td>57</td>
</tr>
<tr>
<td>2.Pick up / drop of</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>3.Leisure</td>
<td>0</td>
<td>0</td>
<td>13</td>
<td>0</td>
<td>0</td>
<td>21</td>
<td>34</td>
<td></td>
</tr>
<tr>
<td>4.Shopping</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>5.Other</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>10</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>6.Home</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>59</td>
<td>13</td>
<td>72</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>41</td>
<td>0</td>
<td>13</td>
<td>6</td>
<td>2</td>
<td>59</td>
<td>82</td>
<td>203</td>
</tr>
</tbody>
</table>

Table 6: Confusion matrix for the trip purpose inference method NB-TPI

<table>
<thead>
<tr>
<th>Observed</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>noRes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.Worktrip</td>
<td>50</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>57</td>
</tr>
<tr>
<td>2.Pick up / drop of</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>3.Leisure</td>
<td>2</td>
<td>0</td>
<td>23</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>34</td>
<td></td>
</tr>
<tr>
<td>4.Shopping</td>
<td>0</td>
<td>0</td>
<td>15</td>
<td>0</td>
<td>1</td>
<td>8</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>5.Other</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>6.Home</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>65</td>
<td>7</td>
<td>72</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>55</td>
<td>0</td>
<td>27</td>
<td>15</td>
<td>6</td>
<td>66</td>
<td>34</td>
<td>203</td>
</tr>
</tbody>
</table>

to be combined with some other method that can infer first visits to POIs, which the remaining TDI methods are able to.

TDI, which is IDW only, achieved the highest accuracy of the static TDI methods that were used to infer public POIs. The accuracy of TDI will not increase when the sample increases. It would instead theoretically reach it’s maximum accuracy when trip definition by the collector is completely accurate and the set of public POIs is complete. When a POI, public or private, is not pre-existing in the database, the user must create that POI. Thereafter, TDI should also be able to infer those POIs at a similar accuracy.

NB-TDI did not perform as well as TDI, but as this is a dynamically learning method and the sample of trips was small, we should expect higher accuracy when the sample increases. This is strengthened by that the accuracy of the method for the second half of the trips analyzed was about twice the accuracy of the first half and close to the accuracy of TDI, although the statistical significance of this division of the small sample of trips in two halves by timestamp is questionable.

The spatial relevancy measure SR-TDI did not achieve higher accuracy than TDI. This shows that, with the parameters and the GIS data set used in this study, this method is not more accurate than IDW. The parameters that was used were: the size of the $R$-neighborhood (100 m) which was where $tf$ was calculated, the size of each cell (10 x 10 m) in the grid where $idf$ was calculated and the number of cells in the grid. A possible
reason for that the performance was not better could be that some POI types in the OSM data are considered less important than others and therefore less well maintained, giving them an inflated spatial relevancy value. Another reason for this could be that there is nothing in SR-TDI that accounts for the general probability of visit to each POI type. This was tested in SRNB-TDI by the addition of the NB classifier but the sample was too small for the NB classifier to build up.

In the full set of trips there were 39 trips where the user created a public POI as the trip destination. This shows that the destination POI was either missing from the OSM data set or was outside the \( R \)-neighborhood. This reduced the number of trips that were used to build the classifier for the NB based TDI-methods and likely also reduced their accuracies. The reasons for the missing POIs were that the OSM data was incomplete in some areas and that the subset selection that was made from the OSM data set left out some important POI types.

Comparing the results of the static and the dynamically learning TDI methods, the dynamically learning method VF-TDI shows promising results in inferring POIs that are revisited. The other dynamically learning TDI methods, NB-TDI and SRNB-TDI, did not have access to enough trips so their performances are unclear. The two static methods, TDI and SR-TDI, performed similar and achieved higher accuracy than NB-TDI and SRNB-TDI with this small sample of trips and are therefore more suitable for short studies. Dynamic methods like NB-TDI and SRNB-TDI might be more useful in longer studies.

Comparing the results of the user specific and the generic TDI methods, the user specific method VF-TDI performed well in inferring revisits to trip destinations while the generic methods TDI, NB-TDI, SR-TDI and SRNB-TDI did not perform as well in inferring public POIs from the OSM data. Generic methods might be more useful when the set of POIs used is has a high degree of completeness.

As the POI type home is almost always associated with the trip purpose home\(^{10}\), and both ITH-TPI and NB-TPI uses the POI type as input, both TPI methods are expected to be accurate in inferring the trip purpose home. This is confirmed as the accuracy of ITH-TPI and NB-TPI were 81.9% and 90.3% respectively. As NB-TPI performed better than ITH-TPI for all each trip purpose using these parameters, the Naïve Bayes Classifier shows to be a promising method for inferring trip purpose. Both TPI methods are dynamically learning, and the generic method, NB-TPI, performed better than the user specific method ITH-TPI.

The set of trip purposes that is used affect the accuracy of the TPI methods. Higher granularity in the trip purposes makes it more difficult to separate each trip purpose, likely reducing the inference accuracy of the methods. None of the TPI methods were able to infer trip purpose pick up / drop of but as the sample of trips annotated with this trip purpose were only 5 trips, the statistical significance of this is small.

A pro of using an extensive set of POIs is that it makes it possible for the TDI methods to infer trip destination without the user having to define POIs manually, reducing the amount of manual work for the users and increasing the chance of automating travel

\(^{10}\)Trip purposes are worktrip, pick up / drop of, leisure, shopping, other and home
surveys. A con is that it requires either higher accuracy in the trip definition, perhaps including indoor positioning that allows for accurate positioning inside buildings, for example inside shopping centers, so that the trip end point is more accurately defined or more sophisticated trip destination inference methods.

Limitations in the web application were identified during the study. The lack of functionality to manually add trips that are not registered by the collector, for example due to battery issues, likely lead to an exclusion of trips. Functionality for a user to create new public POIs where there are POIs missing from the OSM data would likely decrease the number of times where TDI, NB-TDI, SR-TDI and SRNB-TDI does not infer, given that these POIs are visited by several users.

As previous studies did not focus on inferring trip destination, there are no comparable results for the TDI methods. The TPI methods achieved higher accuracies than in the study by Bohte and Maat [1], but as the set of trip purposes are different and the TPI methods in this study use validated trip destinations as input, the results are not directly comparable.

7 Conclusions and Future Work

This thesis set out to explore how two important attributes of a trip, destination and purpose, could be automatically inferred from GPS trajectories and GIS data and has identified methods that show promising results. Previous studies have not used extensive sets of POIs and therefore the density of POIs close to a trip end point would often have been low, which makes trip destination inference less complicated as the inference methods would have less prediction choices. Studies on inference of trip purpose are few and the use of machine learning methods not widely spread. The study sought out to answer the following two questions:

1. Which data and methods are appropriate for trip purpose and destination inference?
2. How can a system based on open-source components that facilitates semi-automatic collection of trips and validated trip attributes be designed?

In summary, this thesis finds that VF-TDI and NB-TPI with POI data from OpenStreetMap are a feasible approach for TDI and TPI, while TDI, NB-TDI SR-TDI, SRNB-TDI and ITH-TPI compare less favorably in terms of the prediction performance. Based on the informal user-experience discussions with the 9 users in the case study and the collected data it was evident that the prototype system with annotation functionality using open source Web GIS technology can indeed effectively facilitate the semi-automatic collection of trips and travel survey data. The remainder of this section is structured so that Section 7.1 gives the conclusions of the thesis and 7.2 outlines suggestions for future work on the subject.


7.1 Conclusions

The suggested dynamically learning user specific method VF-TDI shows promising results in inferring trip destination for trips that were to revisited POIs, reaching an accuracy of 86.8% for revisits to all POIs. The remaining methods that are general and therefore uses all verified trips to OSM POIs achieved an accuracy of 38.9% (TDI), 36.1% (SR-TDI) and 25.0% (NB-TDI, SRNB-TDI), which indicates the need for improvement. Combining VF-TDI with a method that successfully infers first visits to public POIs could lead to an effective trip destination inference method.

The suggested dynamically learning, general method for inferring trip purpose, NB-TPI, shows promising results as it reaches an accuracy of 76.0% on the full set of trips. The fact that it reaches higher accuracy than the dynamically learning, user-specific method, ITH-TPI, where the accuracy was 57.7%, shows that the inference of trip purpose using an inferred / validated trip destination POI and an NB classifier is a promising way of doing this.

The thesis proposes, implements and successfully tests a semi-automated system for the collection of trips and inference / validation of trip destination and purpose. This shows that it is possible to design and implement a fully open-source, web-based trip annotation system that facilitates the collection, validation, inference and annotation of trips in an efficient manner.

Some limitations in the web application likely led to an exclusion of trips. There is no implemented functionality to manually add missing trips, which can be needed if a user’s mobile collector has been turned off or has suffered loss of battery. This might have reduced the number of trips that were used to evaluate the inference methods.

Not all public POIs were present in the data set that was used. The reasons for this were that some POIs were missing from the original OSM data set and that a subset selection of the POIs was performed. There were no functionality set up for users to manually add public POIs and the lack of public POIs reduced the number of trips used to build the NB classifier. This likely reduced the accuracy of NB-TDI and SRNB-TDI, as they are designed to use only public POIs from the OSM data.

7.2 Future Work

The study that was performed in this thesis, where 9 users participated, is considered to be small. To assess the accuracy of the dynamically learning methods it would be interesting to perform a larger scale study that would allow the classifiers to build up, so that the accuracy limits of the methods would be reached when there is a large data sample to use.

It would also be interesting to tune the method parameters. For the TDI methods, the radius around the trip destination, \( R_{\text{dest}} \), was set to 100 m and changing this radius will likely affect the results of the methods. For the methods that include IDW as a component (NB-TDI, SR-TDI and SRNB-TDI) it could be useful to introduce some weighting scheme to study the influence of the IDW component in the methods. For the methods
that use the Naïve Bayes classifier (NB-TDI, SRNB-TDI and NB-TPI) it would be interesting to test different predictor variables to see if they are significant in inferring trip destination and purpose. Some predictor variables that could be studied are trip attributes like travel mode and time at destination, perhaps combined with socio-demographic user variables. For the TPI methods, the set of trip purposes could be reduced or increased which would likely influence the trip purpose inference accuracy.
References


16-06 **Manuela Alvarez.** Mapping forest habitats in protected areas by integrating LiDAR and SPOT Multispectral Data. Master of Science thesis in Geoinformatics. Supervisors: Torbjörn Rost, Metria and Yifang Ban, KTH. June 2016.

