IDENTIFYING THE TRAVEL PATTERNS OF ON-DEMAND TAXI TRIPS THROUGH INFERRED TRIP PURPOSES

Durage Dineth Dhananjaya

218014X

Degree of Master of Science

Department of Transport and Logistics Management

University of Moratuwa

Sri Lanka

SEPTEMBER 2022

IDENTIFYING THE TRAVEL PATTERNS OF ON-DEMAND TAXI TRIPS THROUGH INFERRED TRIP PURPOSES

Durage Dineth Dhananjaya

218014X

Degree of Master of Science

Department of Transport and Logistics Management

University of Moratuwa

Sri Lanka

SEPTEMBER 2022

DECLARATION OF ORIGINALITY

I declare that this is my own work, and this thesis/dissertation does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my thesis/dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: Date: 23/09/2022

UOM Verified Signature

STATEMENT OF THE SUPERVISOR

The above candidate has carried out research for the Degree of Master of Science under my supervision.

Name of the supervisor: Dr. T. Sivakumar

Signature of the Supervisor: Date: 23/09/2022

UOM Verified Signature

ABSTRACT

Activity-based modeling has become the backbone behind transportation planning, and trip purposes that can be inferred from large GPS datasets of different travel means are paving the path to augmenting its accuracy. In this context, the trip purpose inference problem has emerged since the GPS is unable to capture the trip purposes explicitly. This problem has not been thoroughly addressed in developing countries despite the fact that the applicability of a trip purpose inference model extremely depends on the land use context. Hence, this study attempted to ameliorate an accurate model (base model) for a chosen study area in Colombo District, Sri Lanka using the on-demand taxi trips data.

Point of Interest (POI) data is often accompanied by the purpose inference models as it provides a complete insight into the land use around origin and destinations. In the pursuit of the main objective of the study, machine-learning-based text classification was tested to improve the number of informative POIs and its outcome indicated that the Support Vector Machine (SVM) classifier can be utilized effectively and efficiently. The designed trip purpose inference model referring to a base model is proposed as a three-layer trip purpose inference framework in which a method to impute purpose based on trip regularities and the method to identify residential trips was included as two layers before using the base model. The validity of the proposed model was evaluated with the assistance of household travel survey data and with respect to the purpose proportions and division level R-square. Furthermore, travel patterns of the on-demand taxi trips were assessed in terms of temporal regularity, trip lengths, and spatial dynamics. It is recommended to conduct further studies to assess the applicability of other parameters such as trip origin context and trip time with the assistance of unsupervised learning methods.

Keywords:

Travel behavior, trip purpose inference, human mobility, taxi trips, GPS data, POI data, semantic enrichment, developing countries.

ACKNOWLEDGEMENTS

I would like to convey my heartfelt gratitude to Dr. T. Sivakumar for providing me an opportunity to conduct this study under his supervision and for assisting me to drive this towards success from the very beginning. The guidance he has given me assisted to improve the quality of this work gradually and ultimately add a valuable context of knowledge for the community regarding the study area. I would like to extend my appreciation to Senior Professor Amal S. Kumarage for further helping me to identify the sections that required improvements in making this work more worthwhile for both academics and industry.

My warm thankfulness goes to Dr. Amal S. Perera for helping me in data collection and also providing me the additional knowledge from a computer science background which was highly useful during the data processing task. I would also like to thank all the members who worked on the research project “Data-Driven Transport Analytics” and the member of the “DataSearch” research center for the assistance and comments on the work. Furthermore, I would like to convey my sincere gratitude to Accelerating Higher Education Expansion and Development (AHEAD) Operation for financially assisting this study through the above-mentioned project.

Most importantly, none of this would be possible without the love and care of my family. I am grateful to my parents and my brother for strengthening me throughout this study period and helping me to start again with stamina during the critical periods which motivated me to complete this study at level best.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

AFC - Automatic Fare Collection (AFC)

API - Application Programming Interfaces (API)

AGPS - Assisted GPS (AGPS)

AIM - Activity Inference Model (AIM),

CDR - Caller Detail Records (CDR)

DBSCAN - Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

DOP - Drop off Point (DOP)

DMR - Dirichlet-Multinomial Regression (DMR)

DS - Division Secretary (DS)

DT – Decision Tree

KDE - Kernel Density Estimation (KDE)

KNN – K Nearest Neighbors

GPS - Geographical Processing System (GPS)

HG - Human Geography (HG).

HDOP - Horizontal Dilution of Precision (HDOP)

LBS - Location Based Services (LBS)

LR – Logistics Regression

MNB – Multinominal Naïve Bayes

ML - Machine Learning (ML)

NLP - Natural Language Processing (NLP)

NSAT - Number of sync SATellites (NSAT)

OSM - Open Street Map OSM

OSRM - Open Street Routing Machine (OSRM)

POI - Point of Interest (POI)

PUP - Pick Up Point (PUP)

R - Walking Radius (R)

SML - Supervised Machine Learning (SML)

SVM – Support Vector Machine

TRID - Transport Research International Documentation (TRID).

TS - Transportation Science (TS)

UML - Unsupervised Machine Learning (UML)

1. INTRODUCTION

1.1 Background of the Study

Understanding the mobility behaviours of urban communities has become an appealing topic due to the ever-increasing urban population, and the recent pandemic issue also raised its importance. The more complex the urban fabrics become more unpredictable and difficult these behaviours also become. Mobility pattern recognition can help to advance transportation planning, traffic demand management, urban planning, and even increase the usability of location-based services such as mapping platforms. Identifying these patterns on a large scale was a difficult task until recent years despite its advantages, owing to the lack of tools that can be used to monitor individual level time-resolved locations of passenger trips. However, continuous digitization in transportation has created viable means to overcome that barrier. This has been supported by the rapid development of mobile, internet and sensing technologies. Most of these emerged large-scale locational data are technically and economically feasible to use in discovering embedded knowledge.

The utilization of Automatic Fare Collection (AFC) methods such as smart cards for fare collections and the Geographical Processing System (GPS) to track vehicular movements provides large-scale information regarding mobility patterns. Passively collected Caller Detail Records (CDR) are also used in this context. The use of different data sources has various advantages and disadvantages. For instance, CDR data does not have the capability to measure the locations to the level of latitudes and longitudes but at the zonal level. In that case, the recognition of urban mobility patterns is mainly conducted with GPS and AFC data. Although Automatic Fare Collection (AFC) data is often limited to public transportation, GPS data can provide data from either onboard devices of vehicles or wearable devices of passengers. GPS data collection has become even more convenient with the advent of smartphones with their inbuilt GPS sensors. In that case, time-stamped locational data obtained from GPS devices can be considered the most viable approach to use in identifying mobility patterns. In general, GPS data are used to determine the mobility behaviours of individual passengers and passengers who use services such as bike-sharing, car-sharing, and taxis.

Taxi GPS data has become an attractive example within the scientific community among the significant collectible volumes of different GPS-based travel data. In that case, it has been well-recognized that taxi GPS data can be used to extract an in-depth and more holistic understanding of travel behaviours of one of the critical groups of urban commuters. For decades, taxis have played a pivotal role among the available alternative transportation modes for urban communities. The advent of smart mobile devices caused ease in the communication between the passenger and the taxi driver, creating a new means of taxi known as an on-demand taxi. The popularity of this transport mode has increased gradually over the past few years worldwide. Now, a large number of passengers use this service to fulfill their mobility needs. Unlike traditional taxis, on-demand taxi service providers collect GPS data emitted by smart mobiles. Those can be used to analyze in obtaining useful information about the behaviours of both passengers and drivers. Although traditional taxi GPS data are emitted as sequences of time-stamped locations of the taxis' movement (Spatio-temporal points), due to mobile devices, on-demand taxis provide the origin/ Pick Up Point (PUP) and destination/ Drop Off Point (DOP) directly as given by the users. This is a massive advantage in data preprocessing as it consumes a tedious effort to separate the start and endpoints from trajectory data. These data are needed to be enriched to extract the hidden travel patterns more informatively. The purpose of a trip is one of the critical attributes used in this context.

Derived demand of transportation is caused by the requirement of the people to fulfill different spatially dispersed activities. Every trip made by a passenger has a particular purpose or an activity to perform. The activity information about the travel behaviours of passengers is irreplaceable as it is the critical reason behind the generation of travel demand. Thus, it has been used as a predictor to improve the accuracy of the travel forecast models, and it contributes explicitly to identifying the activity-based travel patterns. Hence, enrichment of taxi GPS data with purposes to determine its travel patterns is also complementary to many other advantages. More importantly, the ability to replace the traditional trip activity data collection methods which include many downsides, with a technologically advanced method is a key focus area. However, as raw GPS data does not directly provide any attribute relating to a trip's purpose, it needs to be inferred utilizing a developed methodology. This scenario is known as the trip purpose inference problem.

1.2 The Trip Purpose Inference Problem

The inability of the GPS to identify the activities performed by the people created the trip purpose inference problem, which finds answers to the question of "what did/will a passenger do when arriving at one place?" (Meng et al., 2017). Ermagun et al. (2017) elaborated that this study area has been insufficiently studied compared to the other areas within the transportation domain that utilize GPS data to uncover travel patterns such as mode detection. However, the beginning of this research area goes back around a decade now (Chen et al., 2018). Inferring the purposes of taxi trips is even more challenging as the trip start, and end locations cannot precisely define the venues that may have been visited as generally a passenger walk to the locations after getting off, which the GPS data cannot capture. The problem is built upon an assumption that a passenger performs only one activity during a stop to make the inference possible, although it may not be the case all the time.

The inference of an activity performed by a passenger is difficult to conduct, limited to the data attributes of taxi trips from GPS records. Since a trip is a one-way course to conduct an activity, understanding the locations around the destination is required at the simplest level to infer the trip's purposes. Types of places around the trip destination context provide insight about possible activities to be performed by a passenger. In that case, sort of an understanding of places near the destination context would be an advantage. To this end, Point of Interest (POI) data has been used in most studies, although its accuracy highly impacts the results to be obtained (Li Gong et al., 2016). The inference is performed either by assigning the activity type of the most probable POI that a person may visit or the most probable activity according to the general land use of POIs. Apart from that, other supplementary data such as check-in data of places, road network data, and even survey data have also been utilized by the studies. Nguyen et al. (2020) categorized the existing trip purpose inference methods into three (1) Rule-based, (2) Probabilistic, and (3) Machine-Learning based. The use of specific methods causes the selection of attributes used in a model, which is further elaborated in the following chapters. However, annotation of purposes is generally required at the macro level to identify the travel patterns rather than focusing on the individual level.

1.3 Research Gaps

We identified several critical research gaps to address in this study based on the conducted literature review, which is further elaborated in Chapter 2. This study intends to develop a precise trip purpose inference model to uncover the travel patterns of on-demand taxi trips focusing on the following identified research gaps.

- A comparative assessment of the existing trip purpose inference models to identify the tradeoff between the application has not yet been conducted.
- The impact of POI data quality on the trip purpose inference results and possible enhancement steps has not been addressed despite being highly important.
- Inclusion of trip regularities of the passengers to identify possible trip purposes can be conducted to improve the purpose inference.
- Travel pattern identification from GPS data has not been addressed in developing countries despite having different urban fabrics.

1.4 Research Objectives

The general objective of this research is to develop an accurate trip purpose inference model to identify the travel patterns of on-demand taxi trips. This general objective is achieved by fulfilling the following specific objectives determined from the research gaps.

- To select a more reliable model by comparing a set of widely used existing trip purpose inference models.
- To enhance the quality of POI data for trip purpose inference using machine learning techniques.
- To assess the possible inclusions of trip regularities of the passengers to infer the purposes
- To identify the travel patterns of on-demand taxi trips within a selected study area of Colombo, Sri Lanka.

1.5 Significance of the Study

Chen et al.(2018) mentioned that benefits could be obtained from inferring the trip purposes of taxi trips are mainly for three parties. They are (1) taxi service providers, (2) taxi passengers, and (3) transport and city planners. Taxi service providers can better utilize their vehicles for predicted demand based on different activity types as they are so interested in minimizing the cost of trips to improve their profit margins. This can be further enhanced using activity-based information for real-time ordering of taxis, estimation of fees, etc. One of the most appealing concepts that can be implemented using the predicted trip purpose of taxi trips is the passenger recommendation system. A service provider can use their service as a marketing platform for attractive POIs within a region and provide information such as promotions and new arrival of items for passengers once they take the taxis for similar purposes (Chen et al., 2018). This brings collaborative business opportunities for service providers with other companies. A developed model for taxi GPS data also can be used for different emerging modes such as car rental, bike rental, and even modern sharing options such as bike-sharing and car-sharing (J. Lee et al., 2021). Apart from these advantages, the value of a data attribute for service providers is immeasurable in this era of data analytics and its use of it may come from an unexpected aspect.

The improvement of the service provider side directly impacts to bring indirect benefit for the passengers. The passengers will experience a more streamlined service once the service providers focus on optimizing their operations. For instance, allocating more drivers for morning work trip generation areas helps reduce passenger waiting times. Also, enhanced ride-sharing options with inclusions of activity information cause to reduce the travel cost. The suggested new services such as personal recommendations may improve the usability of passengers.

More importantly, as mentioned at the onset, accurate trip purpose inference helps identify travel patterns, which are crucial to the city and transport planners. The required activity-related data for these parties are traditionally collected using manual travel surveys, which often encounter either under-reporting or over-reporting issues. Also, critical downsides such as increasing cost, longer periods between data collection, and higher nonresponse rates of these traditional methods encourage the use of alternative methods to overcome those barriers. Hence, taxi GPS data as an alternative for a group of commuters makes this requirement more convenient. The city planners can use this collected information from the taxi GPS data to help the government plan urban infrastructure more rationally through efficient land use utilization and better investment decision-making. Activity-based travel demand estimations support transportation planners in planning for future demand in applications such as traffic management, traffic flow identification, new network designing, city logistic route planning, etc. (Luo et al., 2021). These endeavors directly influence the increase in the living standards of the whole society. More specifically, enhancement of aspects such as public safety from better traffic management, and reduction of air pollution from reduced traffic congestion can be achieved by its application. It is worth mentioning that Sri Lanka is also on the verge of replacing the traditional transportation data collection methods with high frequent automatic data methods. Hence, a developed trip purpose inference model suitable for the urban fabric in the country augments this transition.

1.6 Chapter Summary

- It is essential to develop suitable models to identify urban communities' travel/mobility patterns to city and transportation planning due to the ever-increasing urban population.
- Although the trip purpose is a vital attribute, it is not explicitly available for GPS-based trip records for which the trip purpose inference problem attempts to answer.
- A more accurate trip purpose inference model brings different advantages to parties such as (1) taxi service providers, (2) taxi passengers, and (3) transport and city planners.
- To that end, this study attempts to develop an accurate trip purpose inference model to identify the travel patterns of on-demand taxi trips within a selected study area in Colombo, Sri Lanka.
- Since Sri Lanka, as a developing country also on the verge of replacing traditional transportation data collection methods due to the disadvantages, a developed model for GPS-based trip purpose inference can be used to augment that transition.

2. LITERATURE REVIEW

2.1 Overview of the Systematic Literature Reviews

Since the published systematic literature reviews assist in obtaining a clear picture of the subject study area, this subchapter attempts to provide an overview of the already published systematic literature reviews that focus on trip purpose inference using GPS data. Two published journal papers can be found in this context. Gong et al. (2014) reviewed the methodologies published till 2013 on all sorts of trip data that can be derived from GPS data, including trip purposes. Nguyen et al. (2020) specifically reviewed the trip purpose inference studies that use GPS data published till January 2020. Hence, the latter can be identified as the only study directly dedicated to trip-purpose inference (Nguyen et al., 2020).

This systematic review had been carried out on four widely used databases in transport and geography domains and those are ScienceDirect, Web of Science, Scopus, and Transport Research International Documentation (TRID). Four selection criteria that had been used to filter the search results of those databases are the availability of full text, clarity of the methodology, the minimum number of inferred purposes, the minimum number of citations, and 25 publications that had been selected for the review. (Nguyen et al., 2020) categorized the selected studies into two domains: Transportation Science (TS) and Human Geography (HG). This categorization status was widely followed by recently published articles relating to the purpose imputation (J. Lee et al., 2021), (Gao et al., 2021). TS-related studies focus on the movement of people from a place to a place individually, intending to assist the traditional travel surveys directly. The HG-related studies focus on the interrelationship between the movement of people and the environment. Despite having a similar outcome as the results of the studies in these two domains, there are underline differences described in below Table 1.

While the TS-related studies directly become an alternative for the travel diaries, HG-related studies have a scope of identifying the behavior of trips relating to a particular mode of transport such as taxi and bicycles. Hence, it can be determined that the scope of this study from which inference is assessed for taxi GPS data also belongs to the group of HG-related studies. As shown in the table, the number of studies published in this context is relatively less compared to the TS area, which encourages more research to be carried out.

Table 1: Differences between TS and HG-related studies

	Transport Science (TS)	**Human Geography (HG)**
Focus	Deriving purposes individually to assist travel surveys	Enrichment of GPS data to identify aggregated patterns
Data collection	Travel surveys (Primary data)	Probe-vehicle (Secondary data)
GPS device type	GPS wearables	On-board GPS devices
Validation	Ground-truth	Indirectly by previous knowledge
Number of studies	Higher	Medium

***Source:** *(Nguyen et al., 2020)*

More importantly, exploring the types of countries that the trip purpose inference had been carried out in the existing literature indicates that only the applicability for developed countries has been addressed except for China. Because the working, living, transport conditions, and even urban fabrics are different in most developing countries, it is worth focusing on more experimental purpose inference studies in these countries (Nguyen et al., 2020). The following Figure 1 indicates the types of countries of the studies mainly reviewed by the author for this original study.

According to the systematic reviews of (Nguyen et al., 2020) and (Gong et al. (2014), methods that have been used in this context to infer the purposes can be mainly identified as rule-based, probabilistic, and machine learning-based. The methods relating to machine learning can be further categorized into supervised and unsupervised learning. A detailed description of these methods is presented in the following subchapter.

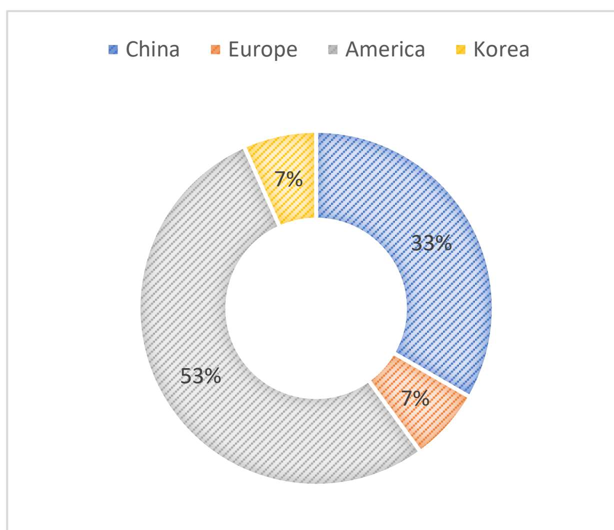

Figure 1: Proportions of Trip purpose inference studies by countries

2.2 Methods of Trip Purpose Inference

2.2.1 Rule-based models

Initially, the trip purpose inference had begun with the deterministic rules as the easiest and most interpretable method (Nguyen et al., 2020). Purposes are assigned to different purposes using assumptions only focusing on the trips' destination context or DOP.

Yue et al. (2012) brought the assumption that all the DOPs around a highly attractive place such as a shopping center can be inferred to the purpose that the POI refers to. A buffer radius is determined to use around the places to capture the DOPs that belong to that place. Despite its simplicity, its use brings difficulties to use in different land-use settings, and most of the urban fabrics contain places belonging to different purposes in proximity.

In contrast to that model, another widely used deterministic rule is the assignment of DOPs to their closest POI (Xie et al., 2009). Unlike providing a radius around the POIs to capture the trips, this method only considers the linear distance between the DOP and POI in assigning the purpose of the closets POI to a trip. This provides an advantage over choosing a purpose for a trip that may have a different purpose for DOPs that are slightly far.

2.2.2 Probabilistic models

Simple deterministic rules bring doubts about their use from different perspectives, such as assigning identical DOPs to different purposes and the influence of other parameters on the purposes rather than simple parameters such as distance, etc. Hence, probabilistic models have been discovered as a remedy for capturing non-determinism. As this approach can confuse spatial and temporal attributes, the transferability to different land-use contexts is higher (Nguyen et al., 2020).

Furletti et al. (2013) proposed a model to infer people's activities from GPS trajectories automatically, and it has been used by many later published studies as a key reference. This study of semantic enrichment of GPS trajectories with activities was conducted as an extension of the works carried out by (Spaccapietra et al., 2008). The activities of users are inferred based on the category of the POI and a probability measure based on the gravity law. POIs are selected only around the DOP, providing a Walking Radius

(R) as a spatial rule and opening times of the POIs as a temporal rule. The numerator of the gravity model is represented by the count of POI that belongs to a specific purpose.

In contrast, the denominator represents the square value of the minimum distance for a POI in that category. The model was tested for a vehicle GPS trajectory dataset in Belgium with available ground truth data for validation and provided an acceptable accuracy rate. The use of spatial and temporal rules in selecting POIs, which is also referred to as candidate POI selection, and the use of gravity law to measure the probability are key follow-ups of this study.

Gong et al., (2016) developed a purpose inference model using the principles of Bayesian probability and considering additional factors such as temporal impact and attractiveness of POIs than the gravity model. Hence, this model can provide the purpose of the most likely activity that a person may perform at different times and zones. Using the Monte Carlo Simulation to include randomness is also suggested in this study. The selection of candidate POIs is performed precisely as Furletti et al. (2013). Due to these two parameters, this model is known for providing more accurate results (S. Gong et al., 2020).

Luo et al. (2021) proposed a hybrid-data approach to infer the purposes using taxi operations data and POI data to estimate trip purposes. Data from a specifically conducted study in Chengdu, China, is used in this study in contrast to other studies in which survey data are neglected to fuse with large-scale trajectory data. The temporal attributes of a trip, including time, day, the date, are embedded in a Bayes model to predict the purposes. Then the spatial attributes are used separately for the prediction. The model provides estimates from the fusion of results obtained from both methods. More specifically, the improvement of the results can be obtained using the POI data highlighted in this study.

2.2.3 Machine Learning (ML) based models

Both rule-based and probabilistic models can infer the purpose by considering the destination context. However, this difficulty can be overcome through the ML-based models by using other parameters such as Origin context. Supervised Machine Learning (SML) models are mainly related to the classification of trips for different purposes for labeled datasets that need to be obtained from ground truth approaches. However, Unsupervised Machine Learning (UML) models are often deployed in uncovering the knowledge behind trips through data mining methods (Nguyen et al., 2020).

Chen et al. (2019) brought up the problem of trip purpose inference as a clustering task consisting of three main components. At first, the neighborhood environment around the PUP and DOP are configured utilizing the POI and check-in data. The use of check-in data makes use of temporal attractiveness of places as different places have different levels of attractiveness according to the time. For instance, an educational POI is far more attractive than another during the morning. The second component provides a more semantic and discriminative context representation by a deep embedding approach. Finally, the K-Means algorithm clusters the records to aggregate the trips with similar latent spaces. However, the use of clustering limited the number of purposes that can be inferred as the optimal number of clusters defines it, and not having a proper validation method is also a specific drawback.

Wang et al.(2017) proposed a generalized probabilistic model defining it as the POI link model for trip purpose inference leveraging POI data at both origin and destination, direction, and trip time. The concept of topic modeling in the Natural Language Processing (NLP) in ML is used in this model. The categories of POIs around the PUP and DOP are considered as documents separately to build the topics. The POI link is constructed, including the PUP and DOP topics pair time. The consideration of intercorrelation between origin and destination is advantageous for this proposed model. The model is further improved to predict the destination and trip time even as proof of the model's predictive power. However, in environments such as the absence of work-related and residential POIs specifically, the application of this model may provide unreliable inferences, which makes it challenging to apply.

Li et al., (2020) designed a model based on the Dirichlet-Multinomial Regression (DMR) topic model to infer the purpose of bicycle trips in Beijing, China. The neighborhood environments around DOPs are defined based on Gong et al. (2016), and it has been used for the DMR model in constructing matrices to annotate trip purposes. The study intended mainly to measure the accessibility of bikes for different purposes, and in that case, the validation was performed primarily by analyzing the spatial distribution of activities.

Gong et al. (2020) proposed a two-layer framework to infer the activities and predict the subsequent trip. The first layer of that model, defined as the Activity Inference Model (AIM), was proposed to reduce the linearity when using the published probabilistic models. To that end, k-means clustering is used separately for the travel distance, DOPs, and drop-off times and annotated the clusters using different heuristic rules, which can be defined according to the nature of the study area. A trip is labeled for a particular purpose if only it had been annotated for the same purpose in three clusters. Otherwise, the model of (Gong et al., 2016) is used to infer the purposes and the Monte Carlo simulation to embed the randomness. This layer of the model enhances the purpose inference, but the selection of rules for cluster allocation for purposes may not provide precise results for some land use contexts.

Table 2 depicts a comparative overview of these illustrated three methods of trip purpose inference in a nutshell.

Table 2: Comparison of trip purpose inference methods

	Rule-based	**Probabilistic**	**Machine-learning**
Transferability	Lower	Medium	Higher
Applicable data scale	Lower	Higher	Higher
Use of parameters	Lower	Medium	Higher
Ground truth data	Optional	Optional	Optional
Interpretation	Higher	Medium	Low
Popularity	Higher	Higher	Medium

2.3 Usage of Different Data Types

As mentioned at the onset, GPS records of human mobility are the primary data source leveraged in this study area. POI data are essential and used in obtaining information regarding the neighborhood environments around the DOPs and PUPs. Apart from those, supplementary data such as check-in data, survey results, and road network data are also used in studies to enhance the outcomes (Chen et al., 2018), (Luo et al., 2021).

Preprocessing and cleaning approaches are often embedded with the use of these data types.

2.3.1 Taxi GPS data

Taxi GPS data are practically obtained from two different sources: (1) devices affixed with the vehicles (S. Gong et al., 2020), and (2) smart mobile phones of the drivers (Merry & Bettinger, 2019). The latter's popularity has increased in the last few years due to the emergence of on-demand taxi services. The inclusion of Assisted GPS (AGPS) systems in mobile GPSs provides accurate records with high sensitivity (Moiseeva & Timmermans, 2010). Simply put, the accuracy of mobile GPSs data is highly reliable with the advent of technologies. GPS data are also collected in two forms: (1) complete trajectory (Zheng et al., 2018), and (2) origin and destination (PUP and DOP) of trips (Li Gong et al., 2016). The use of the former for trip purpose inference consists of a challenging preprocessing step to extract the PUPs and DOPs from complete trajectories (Zheng et al., 2018).

GPS records are often consisting of errors due to factors such as systematic issues with the devices (Wolf, 2006), obstacles caused to signals from buildings and others (Canyoning error), and dispersal of positional data randomly (multipath error) (De Jong & Mensonides, 2003). The most sophisticated devices provide attributes such as Horizontal Dilution of Precision (HDOP) and Number of sync SATellites (NSAT) to assist in removing unreliable records, but not most of the devices available in the market (Tsui & Shalaby, 2006). In the absence of these attributes, the most viable technique that can be used to measure the reliability of the records is to use case-specific measures (Lei Gong et al., 2014). For instance, spatial rules can be used for maximum and minimum allowable distance of trips, and temporal rules can be used for the maximum and minimum allowable travel times of trips to measure reliability (Bao et al., 2017). Records that do not belong to these defined measures are considered erroneous observations.

2.3.2 Point of Interest (POI) data

POI is a specific location point that people may visit to fulfill their different purposes (Pouke et al., 2016). Each POI has a geographical position, a name, a category type it belongs to, and optionally additional information like the opening hours and given user reviews (Furletti et al., 2013). The modern Location Based Services (LBS) systems such as Foursquare (Chen et al., 2019), OpenStreetMap (Pouke et al., 2016), and Google Places (Cui et al., 2018) have made possible options available for users through Application Programming Interfaces (API) to use POI data in different requirements and often accompanied by academics in revealing land use contexts.

POIs with informative category types such as "Bank", "Hospital" etc., are imperative in trip purpose inference. It is the primary reference for a passenger's possible activity visiting that place. For example, by visiting a "Hospital", a passenger generally intends to fulfill a medical-related activity. However, the lack of informative category types of POIs often makes the inference results useless (Furletti et al., 2013). Mostly, these are labeled as "establishment" or generally as "Point of Interest". Another possible implication of utilizing the POI data for purpose inference is the lack of category types relating to specific purposes and work, and residential/home activities are often affected by this (Furletti et al., 2013). Hence, the quality of the POI data is highly influential on the outcomes of trip purpose inference (S. Wang et al., 2019), (Yu et al., 2021). The following Figure 2 depicts the percentage of studies that highlighted the importance of accurate POI data in the corpus of studies referenced in this study.

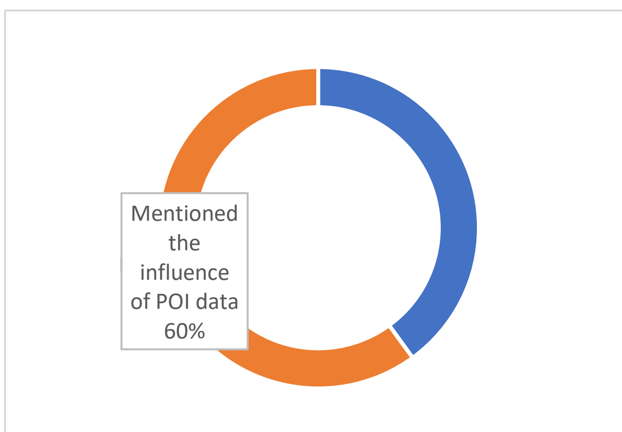

Figure 2: Proportion of studies highlighted the POI importance

2.3.3 Other supplementary data

Check-In data for places are used to augment the purpose inference since POI data only provides the locational reference of different places but not any measure of attractiveness. More specifically, this measures the distribution of different activities commonly taken by passengers in different neighborhoods (Chen et al., 2018). The key advantage of this data type is the ability to measure the attractiveness of different activities in different time frames (Huang et al., 2010). Sources such as Foursquare (Chen et al., 2019) and Twitter data (Chen et al., 2018) are used in this context and often come with preprocessing steps to allocate check-in data to specific POIs. However, unlike the POI data, rich check-in data for many countries and cities are not often available.

Primarily conducted survey data have also been used in studies as a fusion with other data sources (J. Lee et al., 2021). This data type embeds the ground truth data to a certain extent that other sources lack. Use quality survey data can be used to accurately validate the results obtained from purpose inference (Shen & Stopher, 2013). A few studies also accompany road network data to understand better the locations of DOP and PUP from GPS data regarding the road networks. This helps filter the accessible POIs relating to the side of the road around PUPs and DOPs (Chen et al., 2018). The following Table 3 summarizes the different data usage in trip purpose inference.

Table 3: Collecting methods, importance, and accuracy of data for trip purpose inference

Data type	**Collection**	**Importance**	**Accuracy**
Taxi GPS	Affixed devices on vehicles, Smart phones of drivers.	Timestamped locational references of trips	Systematic and random errors may occur.
POI	APIs provided by the LBNS.	Information about places a passenger may visit	Informative categories may miss
Check-In	APIs provided by the LBNS.	Temporal attractiveness of the POIs	preprocessing is required to accurate POI match.
Survey	Author must conduct	Ground truth information	Depend on collection
Road network	From map service providers	Comparative information on the POI and DOP locations with networks	Road type classifications may inaccurate

2.4 Trip Purpose Inference Model Validation

A suitable validation process is imperative for all kinds of model developments to measure the effectiveness of the models. Different kinds of techniques have been used in trip purpose inference studies, and three such approaches are comparing the share of trip purposes (Li Gong et al., 2016), consistency measure with spatial patterns (Sari Aslam et al., 2019), and temporal patterns (Y. Wang et al., 2017), accuracy measure with ground truth data from travel diaries (Furletti et al., 2013) and questionnaires (S. Gong et al., 2020).

However, according to the nature of this study area, validation through a precise accuracy measurement is rather tricky (Nguyen et al., 2020). Comparing with the share of trip purposes is often conducted with pre-conducted general transport surveys that differ from the mode or service provider of the utilized GPS data (Li Gong et al., 2016). Spatial and temporal pattern consistency measurement is rather a qualitative approach that fits with previously known knowledge (Xing et al., 2020). The use of ground truth data also does not match precisely with large-scale datasets due to the small samples. Hence, few studies were even conducted without validation, primarily due to these limitations (Chen et al., 2019). The use of validation methods in the referred corpus of studies is highlighted in Figure 3.

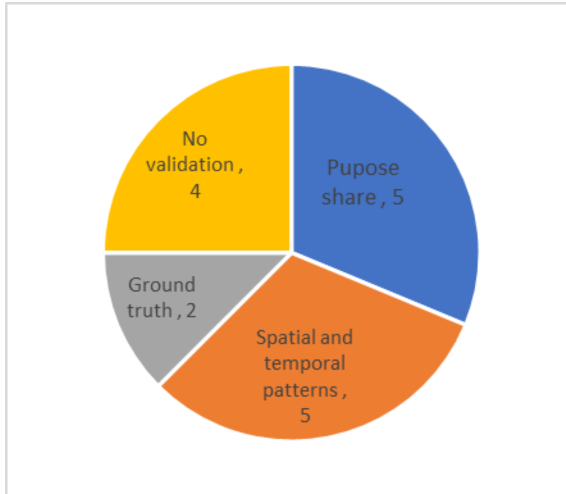

Figure 3: Study count for validation methods

2.5 Travel Pattern Identification

Travel patterns of passengers are crucial in identifying travel behaviors as inputs to transport planners and city planners in decision-making. Many studies that have focused on trip purpose inference made a further step in identifying possible patterns (S. Gong et al., 2020) (Zheng et al., 2018), (Li Gong et al., 2016). Primarily patterns such as activity-based trip lengths, the temporal distribution of trip start and end times, and the spatial distribution of origin and destinations are mainly focused on studies (González et al., 2008). Temporal distributions are often analyzed based on pre-defined time frames, mostly per hour and specifically for weekdays and weekends, and its patterns help deduce decisions. For instance, smoother distribution of activities means passengers tend to make their trips without being bound to time frames (Li Gong et al., 2016). Activity-based spatial distribution of trip origins and destinations is often conducted as spatial clustering or density measurement of defined areas such as grids (Xing et al., 2020). Visualization techniques are imperative to understanding these patterns. Heatmaps and hotspot maps with kernel density estimation are used in this context (Li Gong et al., 2016). The trip length distribution is another important statistic measure that reflects the interaction between people and urban fabrics. Studies have also focused on other patterns, such as frequencies of commuting (Xing et al., 2020) and direction distribution (Li Gong et al., 2016).

2.6 Chapter Summary

- Trip purpose inference in the domain of HG has received lower attention from researchers, and land use contexts of developing countries have not been assessed before according to the best of the authors' knowledge.
- Improving the accuracy of POI data using possible techniques has not been evaluated even though POI data are imperative, and many studies have elaborated its accuracy for the results to be obtained.
- Validation of the developed models is often difficult to be conducted with the mismatch with available ground truth data.
- Activity-based temporal distribution, spatial distribution, and trip length distribution are the most crucial travel patterns that can be analyzed.

3. METHODOLOGY

3.1 Selecting the Study Area

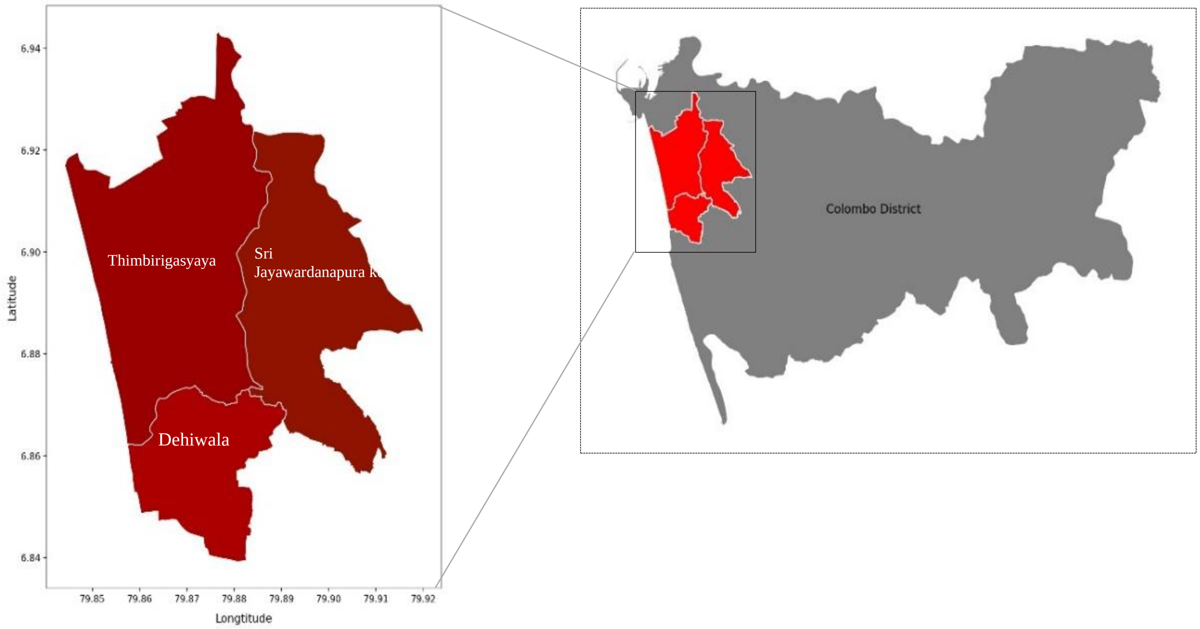

Figure 4: Selected study area for the study

Colombo District of Sri Lanka as the district with the highest population density and the highest average per household income is divided into thirteen Division Secretary (DS) divisions. Among those, ‘Thimbirigasyaya’, ‘Sri Jayewardenepura Kotte’ and ‘Dehiwala’ can be considered as a few divisions that attract a higher number of trips every day due to the urban structure in its’ land-use patterns. Taxi trips that had been attracted to these three divisions from Colombo District through a leading taxi service provider in the country are considered for this study.

3.2 Data Collection and Preprocessing

3.2.1 Taxi GPS data

The primary dataset utilized in this research is the GPS records of the leading on-demand taxi service provider that operates in Sri Lanka. Data were extracted from the data warehouse of the Department of Computer Science and Engineering (DCSE), University of Moratuwa. A Non-Disclosure Agreement (NDA) was signed between the author and DSCE on the use of data for this study purpose while keeping the privacy of the data. Table 4 represents the available attributes and the relevant information about the data.

Table 4: Taxi trip data features information

Field	Example	Description
Trip id	202319411	Unique identifier of a trip
Passenger id	617eac75ef27270524	Unique identifier of a passenger
Driver id	1a07c4dd867254f5c2	Unique identifier of a driver
Taxi model	{1,2,3,4,10,18,24,34,42}	Mode of the trip (i.e., bike, car)
Pick up time	2019-12-24 05:16:43+05:30	Trip book time by the passenger
Actual Pick-up time	2019-12-24 05:20:16+05:30	Trip start time
Drop off time	2019-12-24 05:34:36+05:30	Trip end time
Trip status	{1,2,3,4,5,6,7,8,9,10}	Completion level of a trip (i.e., 2 – cancelled)
Pick up location	7.0006, 79.9497	Spatial reference of trip start location
Drop off location	6.8779, 79.8785	Spatial reference of trip end location
Passenger comments	NA	Given comments by a passenger about a trip (mostly null),
Driver comments	NA	Given comments by a driver about a trip (mostly null),
Driver rating	NA	Rating given by a passenger for the trip (mostly null),

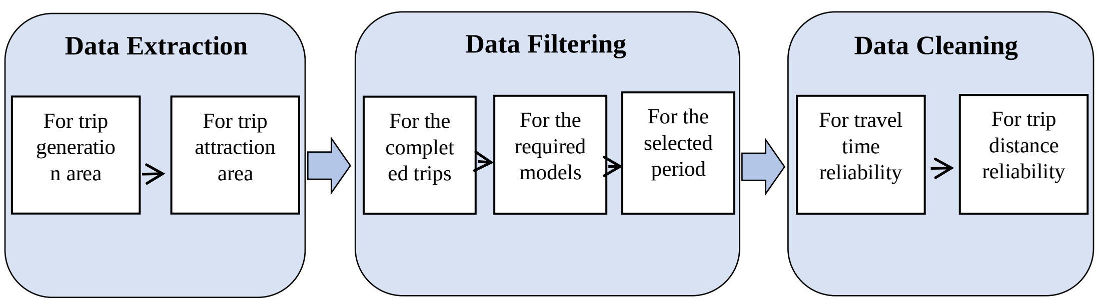

Figure 5: Taxi trip data preprocessing flow

Following the nature of the acquired dataset from the on-demand taxi service provider, data preprocessing should have followed the above depicted three steps in Figure 5. This can be described as a general process to be followed in most scenarios. Figure 7 depicts the spatial distribution of taxi trips across the country. Since data are available for all the operational areas of service providers, it needs to be extracted for the defined study area. In this study, it is the trips generated within the Colombo District and then attracted to the above-mentioned three DS divisions. After the extraction, the data must be filtered for the required trip statuses, selected taxi models, and the selected time frame. Only the completed taxi trips of the three-wheelers, mini car, mini car – 2, and car modes (passenger travel modes) were filtered to be used in the study. Figure 6 shows the available number of trips per mode and the selected set of modes represents more than 95% of the total data.

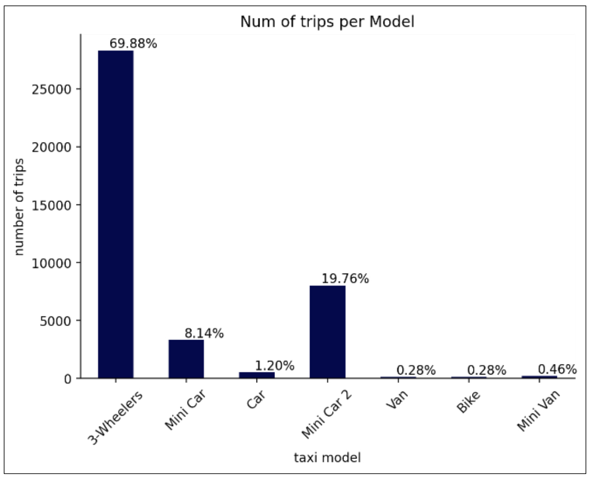

Figure 8: Number of trips by each mode

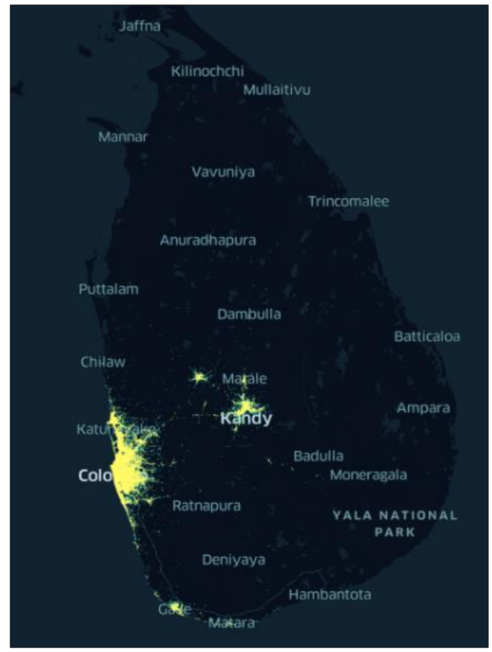

Figure 9: Spatial distribution of taxi trips within the country

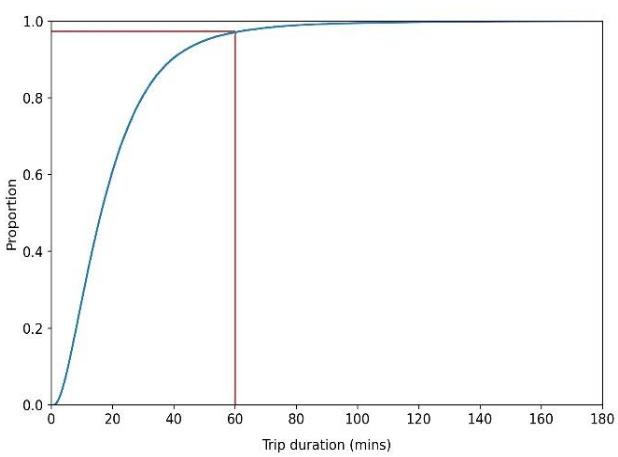

Figure 7: Taxi trip travel time distribution

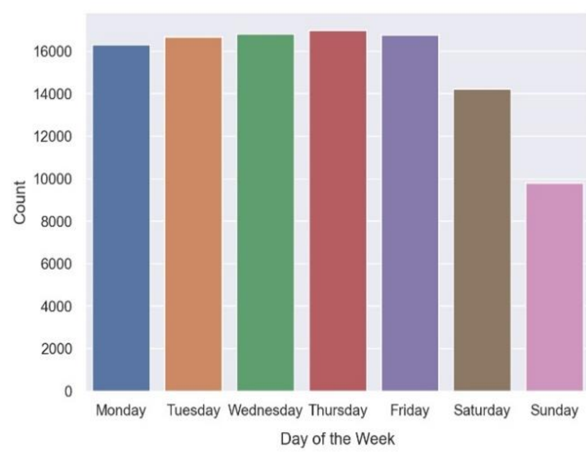

Figure 7: Taxi trip distribution over the period

The period of a week from 11/11/2019 to 17/11/2019 was selected for this study, and then the taxi models relating to the passenger trips and status which represents only the completed passenger trips were filtered. In the absence of attributes provided by the GPS device for error correction, the removal of unreliable records can be done by checking the reliability of travel time and travel distance (Li Gong et al., 2016). For this scenario, it was assumed that a maximum travel time of 3 hours and minimum haversine distance of 200m are potential thresholds to consider a GPS record as a reliable one. These rules were determined based on the generic nature of trips in the study area.

The above Figure 8 depicts the distribution of travel time of the selected taxi trips which indicates that around 97% of trips had been completed within one hour. 107,617 taxi trips (51,637 users) were extracted for purpose inference after the data cleaning process and further illustrated in Figure 9 as per the daily distribution.

The spatial distribution of DOPs of these trips can be seen in Figure 10. The spread of taxi trip DOPs within the selected region can be identified in figure 5. Since 'Sri Jayewardenepura Kotte' has a higher residential land use area compared to the other two, it can be a possible reason for having a lower density of trip attractions. Furthermore, it is visible that a higher number of trips had been attracted along with the route network within this region.

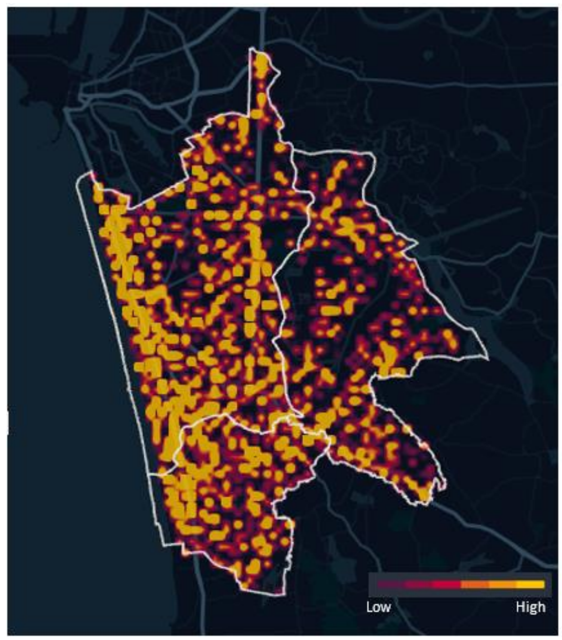

Figure 10: Spatial distribution of selected taxi trips

3.2.2 Point of Interest (POI) data

Google Places API has been used from the available alternatives for most of the recent trip purpose inference studies to extract the POI data because of the data richness (Soares et al., 2019), (Cui et al., 2018). Hence, the "Nearby search request" provided by the Places API was selected to collect the required POI data for the selected study area.

This search request, which is the most suitable among available options, can only collect a maximum of 60 POIs for a given radius around a location. A sample API requests data collection is depicted in Figure 11.

Figure 11: Sample API call

It was decided to collect the POI data for the study area and process it during model building since using the API requests iteratively is cost-ineffective. For that, an algorithm was developed to call the API requests and collect the POI data for a square(grid) region. The selected study area also was divided into a set of square regions and once the data for all the squares had been collected, data relating to the study area was filtered.

This grid-based mechanism was followed in data collection for several reasons. Since the radius of the API call cannot be larger due to the limitation of maximum POI extraction it has to be an optimal value which in this case it is used as 50m. Iteration of circular areas leaves empty spaces in between which only can be overcome by calling the API from the center of the space shown in Figure 12. Also, the use of larger square regions for the whole study area tends to reach a larger unnecessary area which costs the data collection. It is worthy to mention that a marginal gap around the boundaries of the selected study area was given to ensure that DOPs at the boundary may not miss the POIs that a passenger may visit.

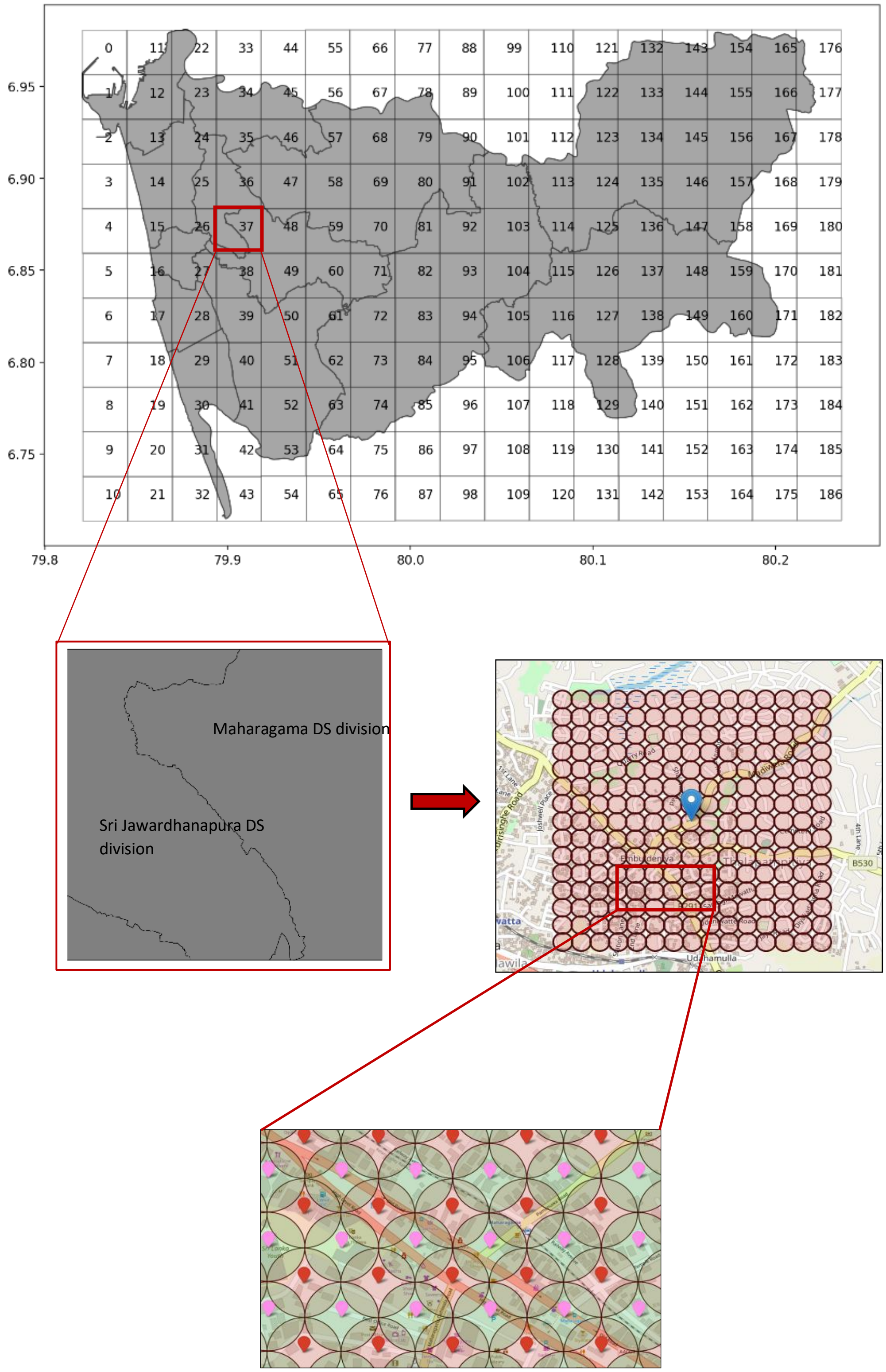

Figure 12: Grid-based POI data extraction approach

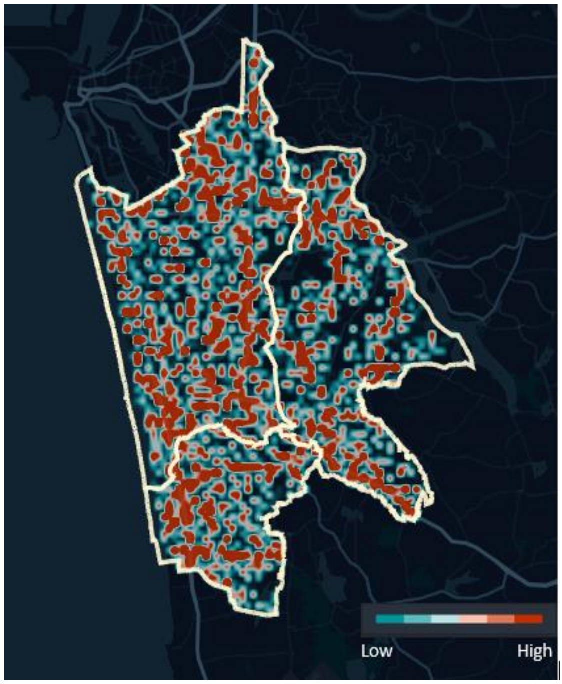

Figure 13: POI data spatial distribution

A total number of 42,418 POIs were extracted to use for this study. The spatial distribution of the POIs for the study area is shown in Figure 13. As crowdsourced data are often embedded with the data cleaning approaches, there is no exception for the POI data as well. The initial issue of POI data was the allocated multiple category types for a single POI. The most suitable category type was assigned manually by identifying a few embedded patterns to avoid multiple category selection.

Collected POIs had been labeled into 97 category types and were further allocated to 8 categories for purpose inference as "Personal, "Educational", "Shopping", "Medical", "Recreational", "Dining" "Transit" and "Residential" as shown in Table 5. Most of the existing trip purpose inference studies also followed a similar level of categorization (Li Gong et al., 2016), (Furletti et al., 2013). However, 15,483 (36.5%) of the POIs did not have an accurate category label. A method to overcome this issue is addressed in section 3.4.1. Also, "Work-related" trips do not belong to one in those seven despites being one of the key trip purposes among urban communities, as a specific category type could not be assigned according to the categorization of the API.

3.2.3 Route network data

Geospatial route network data with route type classification was required as supplementary data to measure the distance between POIs and roads in the developed methods to enhance purpose inference and identify the trips attracted to residential places. OpenStreetMap (OSM) which is a crowdsourced geographic database, is often regarded

Table 5: POI categorization

Purpose	Category Types					
-	store	clothing store	electronics store	home goods store	grocery	shopping mall
Shopping	pet store	shoe store	department store	convenience store	movie rental	campground
-	supermarket	jewelry store	hardware store	furniture_store	bicycle store	book_store
Education	school	university	library	secondary school	primary school	
Medical	health	hospital	doctor	pharmacy	dentist	physiotherapist
-	drugstore					
Recreational	park	night club	museum	art gallery	tourist attraction	movie_theater
-	zoo	casino	aquarium	stadium		
-	travel agency	beauty salon	car repair	place of worship	bank	atm
-	car wash	mosque	post office	hindu temple	finance	funeral home
-	car dealer	real estate	lawyer	cemetery	police	insurance agency
Personal	plumber	locksmith	painter	city hall	car rental	accounting
-	lodging	establishment	apartment	moving company	courthouse	natural feature
-	gym	government office	spa	veterinary care	synagogue	florist
-	electrician	fire station	premise	roofing contractor	storage	general contractor
-	church	hair care	laundry	embassy		
Transport Transfer	transit station	bus station	train station	taxi stand		
Dining	restaurant	food	bakery	cafe	meal takeaway	meal delivery
-	liquor store	bar				

as the best national-level source for such data for many countries (Barrington-Leigh & Millard-Ball, 2017). Therefore, OSM of the study area was processed to extract the required information in this study. Few inaccurate route type classifications were corrected manually by assigning the accurate type.

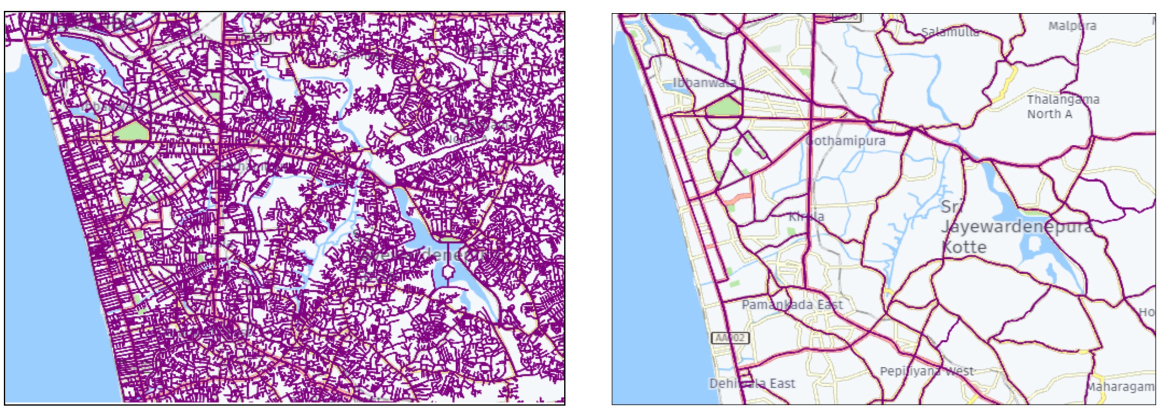

Figure 14: OSM Road network data (left: all roads, right: main roads)

3.2.4 Household travel survey data

Another supplementary dataset was required for the analysis conducted to identify the purpose of regular taxi trips (section 3.6) and validation of the final developed model (section 3.8). The most suitable available dataset for the selected study area is the published dataset collected under the Urban Transport System Development Project for

Colombo Metropolitan Regions and Suburbs. This project, which is also referred to as CoMTrans, is intended to develop an efficient, safe, and reliable transport network for the Western Province of Sri Lanka. One key objective of it was to collect reliable transport data to be utilized in development projects scientifically through surveys. A household travel survey had been conducted in 2013 in fulfillment of this objective. This data is utilized for the above-mentioned requirement and the major downside of it is the impact that occurs from the data collection period difference between survey data and GPS data.

3.3 Selecting the trip purpose inference models to compare

Rule-based and probabilistic methods that focus on the destination context (around DOP) of a trip are the most widely used among the three trip purpose inference methods (Nguyen et al., 2020). Hence, the main three trip purpose inference models from those two methods were selected to comparatively assess the applicability for the selected study area and to choose the most suitable one. A common scenario of considering the destination context for purpose inference is to select the candidate POIs for a given DOP.

3.3.1 Selecting the candidate POIs for a DOP

The activity to be performed by a passenger once got off of the taxi should be within proximity relative to the DOP. Thus, only the POIs within an allowable walking radius must be considered in the purpose inference for a given taxi trip. Density in the distribution of POIs within the considered land use context mostly impacts determining the walking radius because a sufficient number of POIs should have been captured for inference (Furletti et al., 2013). Hence, this value varied from 100m to even around 1000m in the previous studies (Burbidge & Goulias, 2009), (Li Gong et al., 2016). The walking radius is determined as 100m for this study since the selected study area has a highly dense spatial POI distribution. POIs within the walking radius are eligible to become candidate POIs in the purpose inference for a given taxi DOP. To ensure that a POI is a candidate POI, it must also be within the opening hours for the timestamp of the DOP. Figure 15 highlights this scenario of selecting candidate POIs from the destination context for a taxi trip based on its DOP.

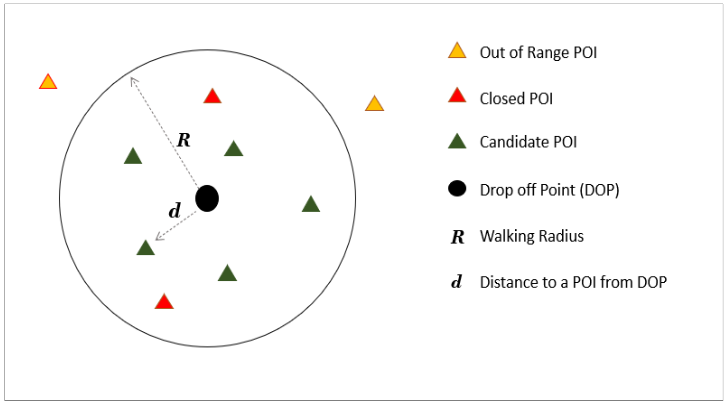

Figure 15: Candidate POI selection

3.3.2 Selected trip purpose inference models to compare

The selected first model can be defined as the closest assignment, which simply assigns the closest POI for a given DOP. It assumes that a passenger is likely to get off from a taxi at the closest proximity to the place where the activity is performed.

$$Trip\ purpose = Category\ of\ the\ Closest\ POI$$

Equation 1: Closest Assignment

The second is the gravity model introduced by Furletti et al., (2013), which considers both the distance for a POI from DOP and the number of POIs belong to different purposes within the walking radius, as shown in Equation 2.

$$P(Ai) \propto \frac{1 * (Count\ of\ POIs\ for\ purpose\ type\ i)}{(Minimum\ distance\ from\ DOP\ to\ a\ POI\ of\ purpose\ type\ i)^{2}}$$

Equation 2: Gravity model

The third model is the one introduced by L. Gong et al., (2016) based on the Bayes theorem and it is depicted below Equation 3.

$$P(O_i\,|(x,y),t) = \frac{P\big((x,y)\big|O_i\,,t\big) * P(O_i\,|,t) * P(t)}{P((x,y),t)}$$

$$P\big((x,y)\big|O_i\,,t\big) \propto A_i * d((x,y),O_i\,)^{-\beta}$$

Equation 3: Bayes-theorem based model

Visiting Probability for a POI (O_i) given the DOP location (x, y) and time (t) is calculated through this model. It assumes that DOP location and its timestamp are conditionally independent given the POI (Li Gong et al., 2016).Attractiveness value (A_i) and distance from DOP to POI (d) are used to estimate the $P((x,y)|O_i\,,t)$ as depicted in (4), which is influenced by the gravity model. A_i is an assigned discrete value that should be based on the generic trip attraction nature within the considered area. Hence, in this study the attractiveness was given as recreational < dining < transit < medical <education <shopping <personal <work, according to the obtained insights from taxi passengers. The distance decay parameter (β) was selected to be consistent as 1.5 following the results from existing studies (Li Gong et al., 2016). $P(O_i\,|,t)$ as the probability to visit a particular POI (O_i) given the time (t) refers to the temporal distribution of activities. It was determined by L. Gong et al., (2016). referencing the output from (Burbidge & Goulias, 2009), which provides a matrix of discrete values for every hour regarding the different purposes. Hence, a similar kind of matrix was created for the study area for Weekdays, Saturdays, and Sundays separately. Thus, the exact equation to be utilized under this model in purpose inference can be simplified to the following Equation 4.

$$P(O_i\,|(x,y),t) \propto \frac{A_i * d\big((x,y),O_i\big)^{-1.5} * P(O_i\,|,t)}{\sum_{i=1}^{n} A_i * d\big((x,y),O_i\big)^{-1.5} * P(O_i\,|,t)}$$

Equation 4: Explicit equation for bayes-theorem based model

3.3.3 Crowdsourced reviews to measure the POIs' attractiveness.

The Places API provides the number of reviews along with the POI data, counting the reviews given by users of the map service. Though it cannot capture the temporal variations of attraction as check-in data, it may provide an advantage over discrete values in finding the place a person most likely goes to. Two difficulties needed to overcome in this evaluation were that some POIs may not have at least a single review while some may have substantially higher values that could go beyond even 17,000. Suppose the number of reviews is to be used directly as the attractiveness value. In that case, some POIs may be dropped having a zero value and some may be assigned directly having an ample value for attractiveness. Equation 6 was designed to overcome these two complications.

$$A_i = \ln(number\ of\ reviews_i + e)$$

Equation 5: Designed equation for attractiveness measure

The usage of the logarithm function concise the distribution of the number of reviews and the addition of Euler's number (e) helps to take the minimum value of A_i as 1 when no reviews were given for a POI. Following Figure 16 highlights the distribution for the number of reviews obtained from places API and the corresponding attractiveness values determined from Equation 5. The model obtained using this factor is defined as the updated Bayes model hereafter.

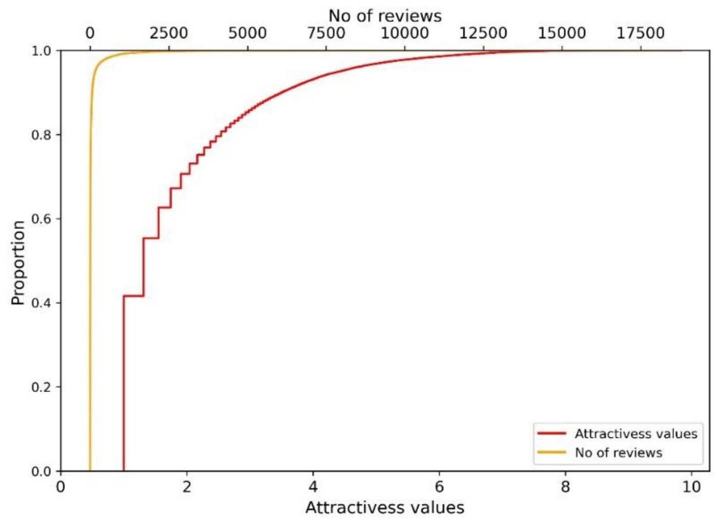

Figure 16: Attractiveness value estimation from number of reviews

3.4.5 Assessing the temporal impact for trip purposes within the study area

The denoted Bayes-theorem-based model from equation 4 consists of a parameter to provide the probability to visit a particular POI (O_i) given the time (t) as mentioned. It was imperative to replace the values used by the original study to apply them to the local context. Hence, a survey was conducted to determine the values to be used for this study following the values used in the original study.

The sample size required for the survey was determined by using 90% as the confidence level and 10% as the margin of error. This resulted in a sample size of 68 about the population size (51, 637 total users). Passengers who travel by on-demand taxi services within the selected study area were the targeted group to obtain the responses. The original study used discrete values for each purpose to determine this required probability (Burbidge & Goulias, 2009). Hence, the conducted survey was designed to obtain the outcome the same which demands the most probable activity that a respondent may perform in specific periods. The main tool used for the survey was Google Forms and distributed online to obtain the responses. A sample survey form is provided in Appendix 1, and Appendix 2 indicates the used results for this study.

3.4 Enhancing the POI Data for Trip Purpose Inference

3.4.1 Predicting the category of a POI

One of the most important attributes of POI data for the purpose of inference is the category that a POI belongs. However, POI data that can be extracted from different sources mostly lack the category type for some instances, which makes the inference results useless. In section 3.2.2 it was mentioned that the POI dataset of this study lacks precise category types for 45% of the data. Intuitively, this issue can be solved using text classification as a model can be built to predict the category type based on the name of that POI. Simply put, when a POI has a name as "National Primary School", category of it can be predicted as educational.

K-Nearest Neighbors (KNN), Logistics Regression (LR), Random Forest (RF), Multinomial Naive Bayes (MNB), and Support Vector Machine (SVM) were evaluated to predict the category type of a POI based on the name. The classification task was followed by a Bag-of-Words model after lowercasing and removing the unnecessary characters. Words with only one occurrence were removed as it does not impact the model performance but reduce the processing time. This resulted in 5,242 words to be used as features for the classification. 80% of the dataset (18,516 instances) was used for training while 20% (4629 instances) was set aside for testing. F1-score as shown in Equation 6 was used to evaluate the model performances and the results are depicted in Table 6. SVM which was trained using the stochastic gradient descent optimization technique can be selected as the suitable model concerning the obtained F1 scores.

$$Precision = \frac{True\ positive}{True\ positive\ + False\ positive}$$

$$Recall = \frac{True\ positive}{True\ positive\ + False\ negative}$$

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Equation 6: F1 score to evaluate ML classifiers

Table 6: ML classifiers results for POI category prediction

Model	**KNN**	**LR**	**RF**	**MNB**	**SVM**
Model Accuracy	0.641	0.694	0.735	0.745	0.761
F1 Scores					
Personal	0.64	0.65	0.74	0.74	0.76
Educational	0.69	0.75	0.78	0.80	0.81
Residential	0.85	0.85	0.89	0.87	0.88
Shopping	0.62	0.78	0.70	0.72	0.72
Medical	0.56	0.70	0.72	0.77	0.78
Dining	0.61	0.67	0.74	0.75	0.75
Transport Transfer	0.81	0.82	0.85	0.90	0.82
Recreational	0.24	0.44	0.36	0.20	0.43

This is often regarded as one of the best models for the text classification problem (IKONOMAKIS et al., 2005). Hence, it was decided to use an SVM classifier to label the POIs without accurate categories. The lower accuracies of the recreational purpose are subjected to the lower support of the utilized test sample. However, categories with lower supports were not dropped in model building as a complete categorization assists to improve the purpose inference results better than an incomplete categorization.

3.4.2 Identifying the Entrance Locations of POIs

There are POIs which are highly attractive but located far from the roads due to the large land areas of those such as schools, conference halls and hospitals. This becomes an issue in purpose inference as the given location does not fall within the provided Walking Radius (R) around the DOP of a trip (section 3.3.1). Identifying the entrance locations of these POIs helps to reduce this issue. GPS data of human mobility is required for this, and taxi GPS data are used in this study to that end. Clustered GPS data of trip attractions with possible signs of entering specific locations can be used to overcome this issue, although it is required a complete method with assumptions to apply.

The following Figure 17 depicts the total process for identifying the entering locations of POIs. The proposed solution to overcome this issue is followed by a prerequisite of selecting the POIs that require entering locations to improve purpose inference. It is assumed that a POI should have a higher attraction level and a considerable distance from the nearest road to get selected. This causes to select of only impactful POIs that are neglected through the allowed maximum Walking Radius (R) around the DOP of a trip. The two assumptions are assessed by using the 'number of reviews' attribute provided with the Places API and measured 'linear distance from the nearest road' respectively. Use of k-means clustering to these two features and selecting the clusters with a higher number of reviews and higher distances to assist in selecting the required POIs. The advantage of using a clustering method for this purpose comes with its capability to group similar data and provide the ability to select the required groups.

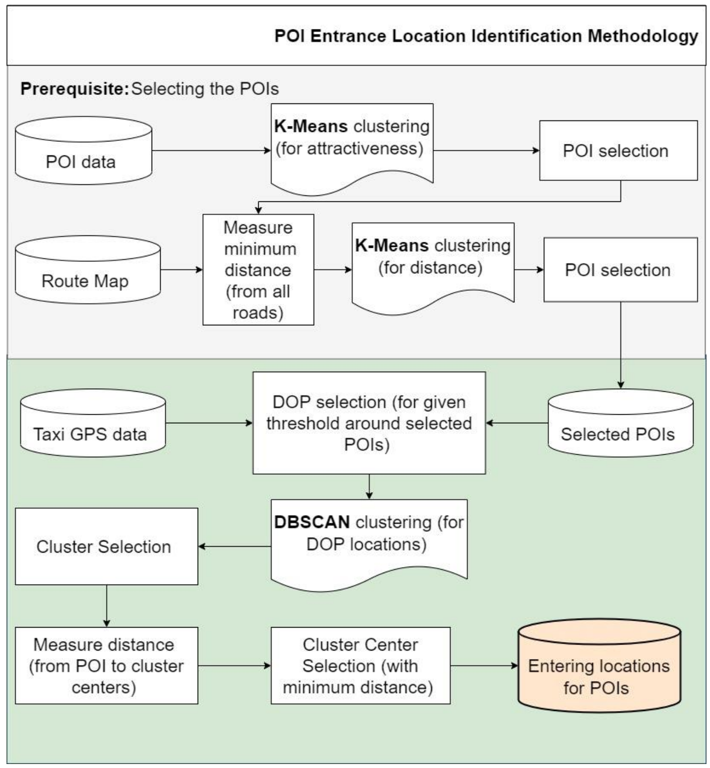

Figure 17: POI entrance location identification method

Providing a Radius (R) for a POI assist to select the taxi trips that had attracted to the surrounding of it using the taxi DOPs. Then, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm can be used with its optimized parameters for the DOPs to capture the clustered trip attraction points around that POI (Ester et al., 1996). Two parameters such as maximum distance between two DOPs and minimum number of DOPS that need to select the clusters are required to optimize the algorithm. This assists to select the possible entering locations by taking the center of the cluster with the minimum distance from the POI.

3.5 Identifying the Trips Attracted to Residential Places

The POI dataset does not contain the general places such as houses of residents apart from apartments. However, neglecting this causes to infer the trips attracted to residential places to other POIs. This issue was required to minimize in order to obtain results with higher accuracy.

Three assumptions are used in this developed approach to identify the trips attracted to residential places. First is for a given day, the DOPs cannot exist with clusters as the residential places are the least attractive compared to all the other types of POIs. Secondly, the residential places do not exist very close to the Main roads (Expressways, Major and Minor arterials), but to the local streets. Third, if a trip was intended to a normal POI along the local streets, the passenger had got off at very close proximity to that POI. This is because the POIs along the local streets usually have a parking area and also the traffic conditions of local streets are not high to compel a taxi to drop a passenger from a different location.

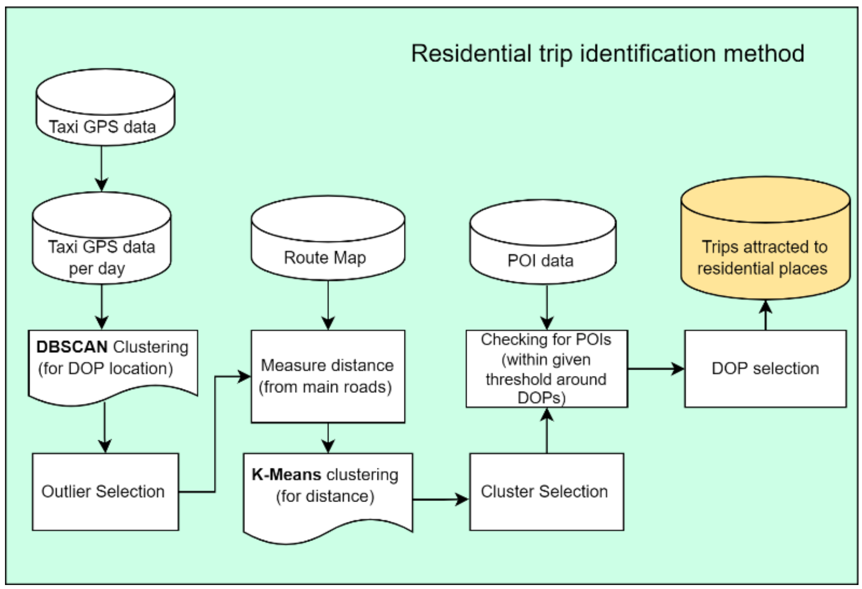

Figure 18: Residential trips identification method

Figure 18 shows the steps that should follow to select the taxi trips attracted to residential places based on the given three assumptions. The main concern is that the data should process day by day relating to the first assumption. Thus, DOPs apart from the clusters can be selected from the outliers of an optimized DBSCAN algorithm. Then the use of k-means clustering to the distance for a taxi DOP from the closest main road and removal of clusters with lower distances can be used for the second assumption. A specific threshold value for this approach could not be used in this study to exclude the clusters due to the non-availability of such parameters. However, the utilization of such parameters may assist in obtaining highly reliable results. The maximum distance from the main road of each cluster assists to select the required clusters. Finally, if there is a POI within a given very narrow distance threshold for the DOP it can be considered as a possible trip that is attracted to that POI. Taxi trips that come out through this pipeline are the selected trips that were attracted to the residential places.

3.6 Identifying the Purposes from Trip Regularity Analysis

Passengers can be divided into commuters and non-commuters. Commuters repeatedly travel for specific purposes, mostly such as work-related and educational. These trips are regular in both space and time (S. Lee & Jang, 2019). Hence, this scenario can be utilized to infer the purposes of regular passenger trips.

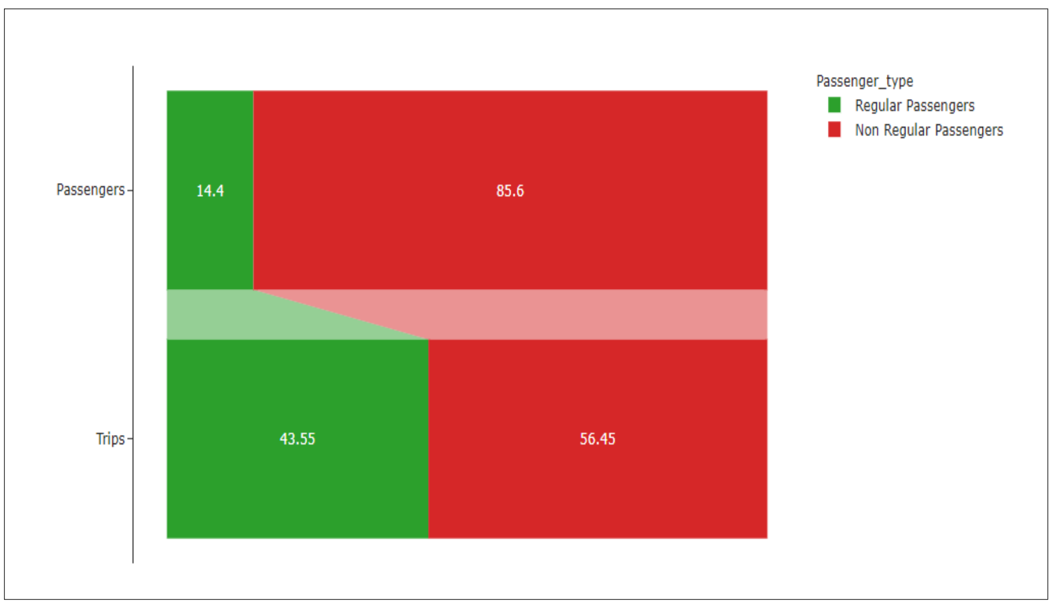

Figure 19: Trip distribution between commuters and non-commuters

Figure 19 was taken by considering the passengers who made more than one trip per week as regular passengers, and it indicates that despite having a lower proportion of passengers the number of trips made is considerably higher.

3.6.1 Scenario identification from spatial regularities

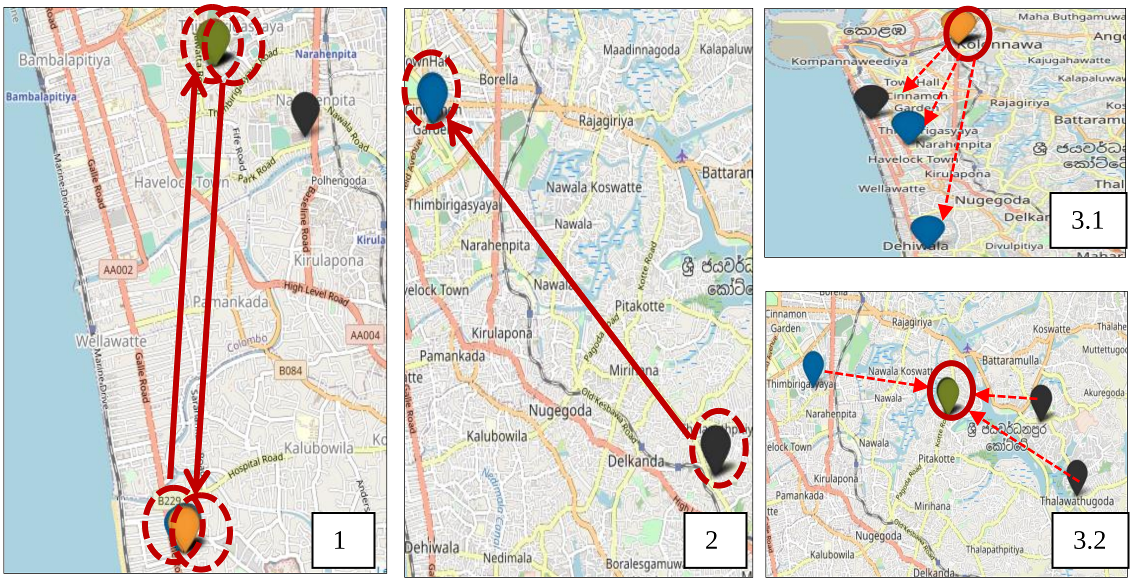

Figure 20: Identified regular trip scenarios. (1) Round trip, (2) Onward trip, (3.1) Regular PUP, (3.2) Regular DOP

In this context, spatial and temporal regularities are accompanied by commuters. Spatial clustering was conducted at both PUP and DOP locations as per passenger at first to identify possible scenarios for purpose inference since it is more important than temporal regularity from an analytical perspective. This was performed using the DBSCAN algorithm and the identified main three scenarios are shown in the above Figure 20 and described in Table 7.

Table 7: Descriptions for identified regular trip scenarios

	Scenario	Description
1	Regular round trip	2 PUP clusters + 2 DOP clusters (1st PUP cluster location = 2nd DOP cluster location)
2	Regular onward trip	1 PUP cluster + 1 DOP cluster
3 (a)	Regular PUP location	1 PUP cluster
3 (b)	Regular DOP location	1 DOP cluster

In general, the regular trips made by the commuters are often home-based and should contain at least one PUP cluster and one DOP cluster. Hence, it was decided to use only the 1st and 2nd scenarios to identify the purposes of regular trips. Work-related and educational are generally fulfilled by regular trips and were selected to assign from these two scenarios.

3.6.2 Selecting the time bins for trip purposes

Having selected two scenarios of trip regularities from spatial clustering, it is required to consider the temporal regularities and an approach to differentiate the trips into the selected two purposes. The concept of time-bin allocation helps to achieve both requirements to a certain extent (Goel, 2018). Household travel survey data is used to identify the peak travel times of work-related and educational trips attract to the selected study area as shown in Figure 21. Two-time bins, “6:00 – 9:00 a.m.” and “12:00 14:00 p.m”, were chosen for work-related trips, and “6:00 – 8:00 a.m.” time bin was chosen for the educational trips as those encounters for more than 90% of the trips recorded in survey data.

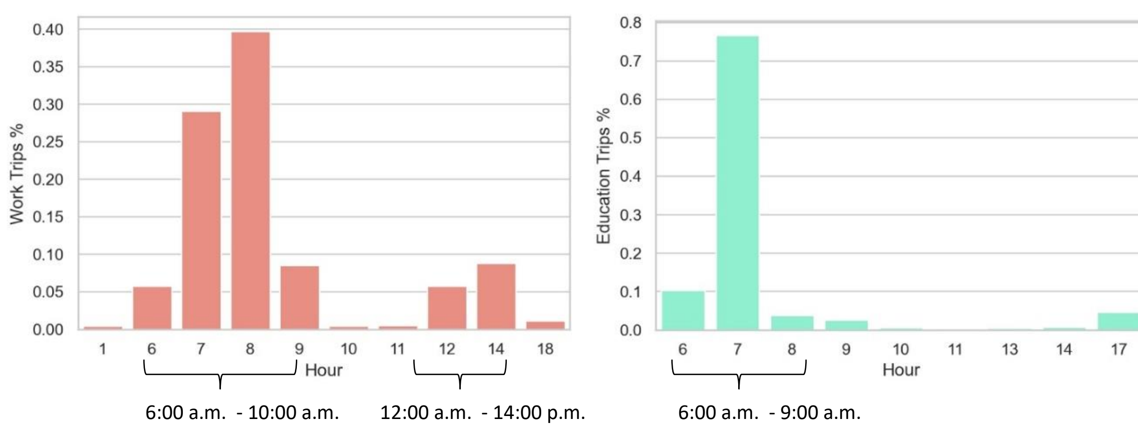

Figure 21: Time bin selection for regular trips

The overlapping time bins for work-related and educational purposes (“6:00 – 9:00 a.m.” and “6:00 – 8:00 a.m.”) doubts the assigning of regular trips of a passenger to a one. This complication can be solved by considering the POIs around the DOP cluster.

3.6.3 Purpose imputation of the round trips and onward trips

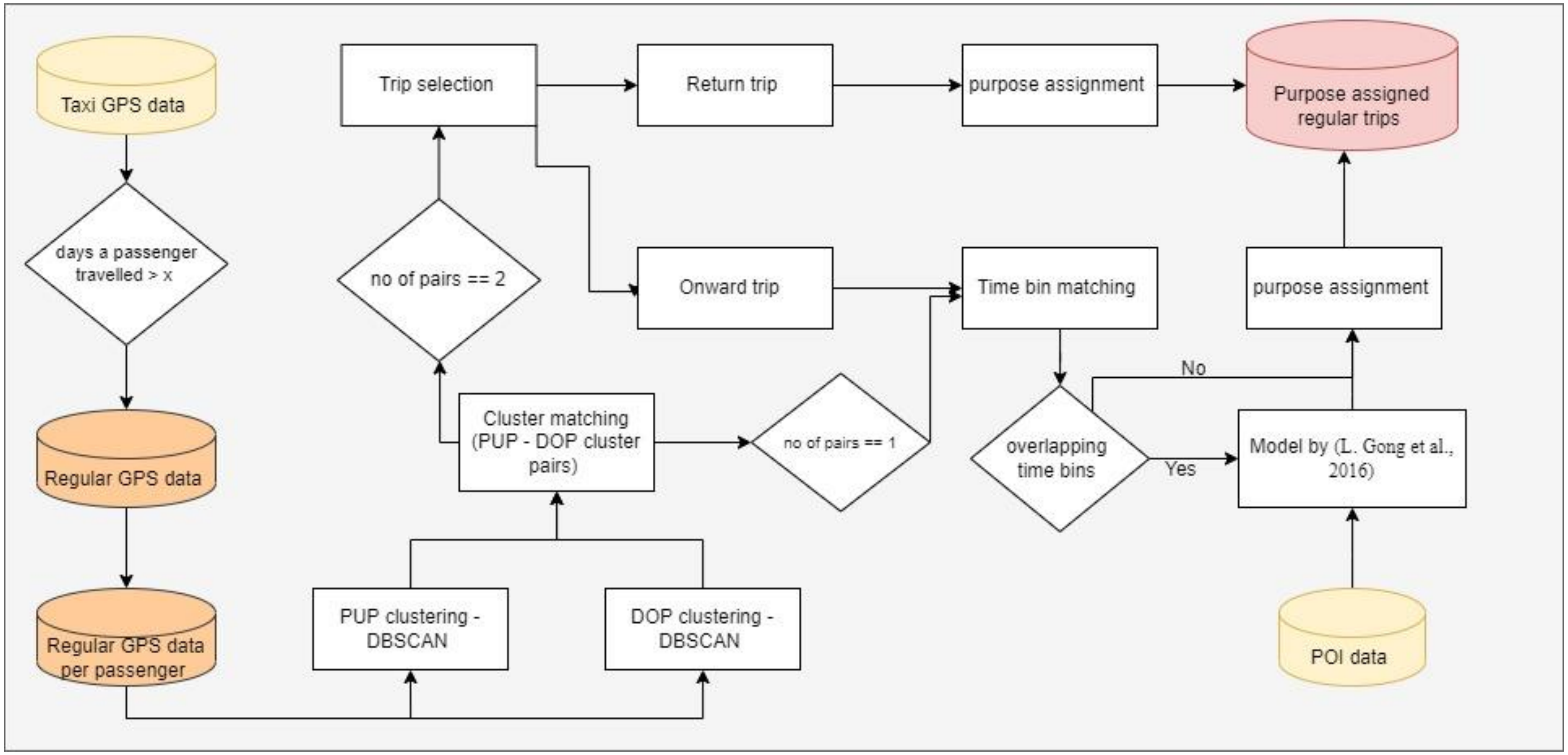

Figure 22: Steps of the designed purpose imputation method through trip regularities

If educational POIs exist around the DOP cluster of a passenger, and it matches with a rational purpose inference model, then it can be assigned to the educational purpose over the work-related despite having similar time bins. The model of (Li Gong et al., 2016) is used for this requirement, in this study.

The complete process of trip purpose imputation through round trips and onward trips is shown in above Figure 22. Several critical facts are addressed in this diagram:

- Trip purpose identification from regularities is needed to be analyzed per identified regular passenger separately, and for all the passengers iteratively.
- Since these trips are assumed to be home-based trips, return trips of the round trips are considered trips attracted to residential places.
- Temporal regularities are addressed by matching the Drop off Times (DOTs) of the identified DOP clusters with the defined time bins (time bin matching).

3.7 Designed Three-Layer Framework for Trip Purpose Inference

A complete trip purpose inference framework was developed as the outcome of this study addressing the achieved objectives. It is expected to provide a convenient reproducibility of the conducted set of works to improve the domain of trip purpose inference through this complete framework. This framework consists of three subsequent layers, each focusing on imputing the purposes for a set of pre-collected taxi GPS datasets.

- Layer 1:
    - Impute the purposes for identified round and onward regular trips (based on the constructed method in section 3.6.3)
    - Imputation through regularities is generally more reliable with the minimum usage of POI data as errors in POI data don't impact on large.
    - This layer is essential as POI-based inference does not support work-related trips, due to the absence of labels.
    - Parameter of "days a passenger traveled", which means the minimum number of days a passenger should travel to consider as a regular passenger, is required to be determined. This can be done by comparing the work-related trip percentage for the study area from the household travel survey.

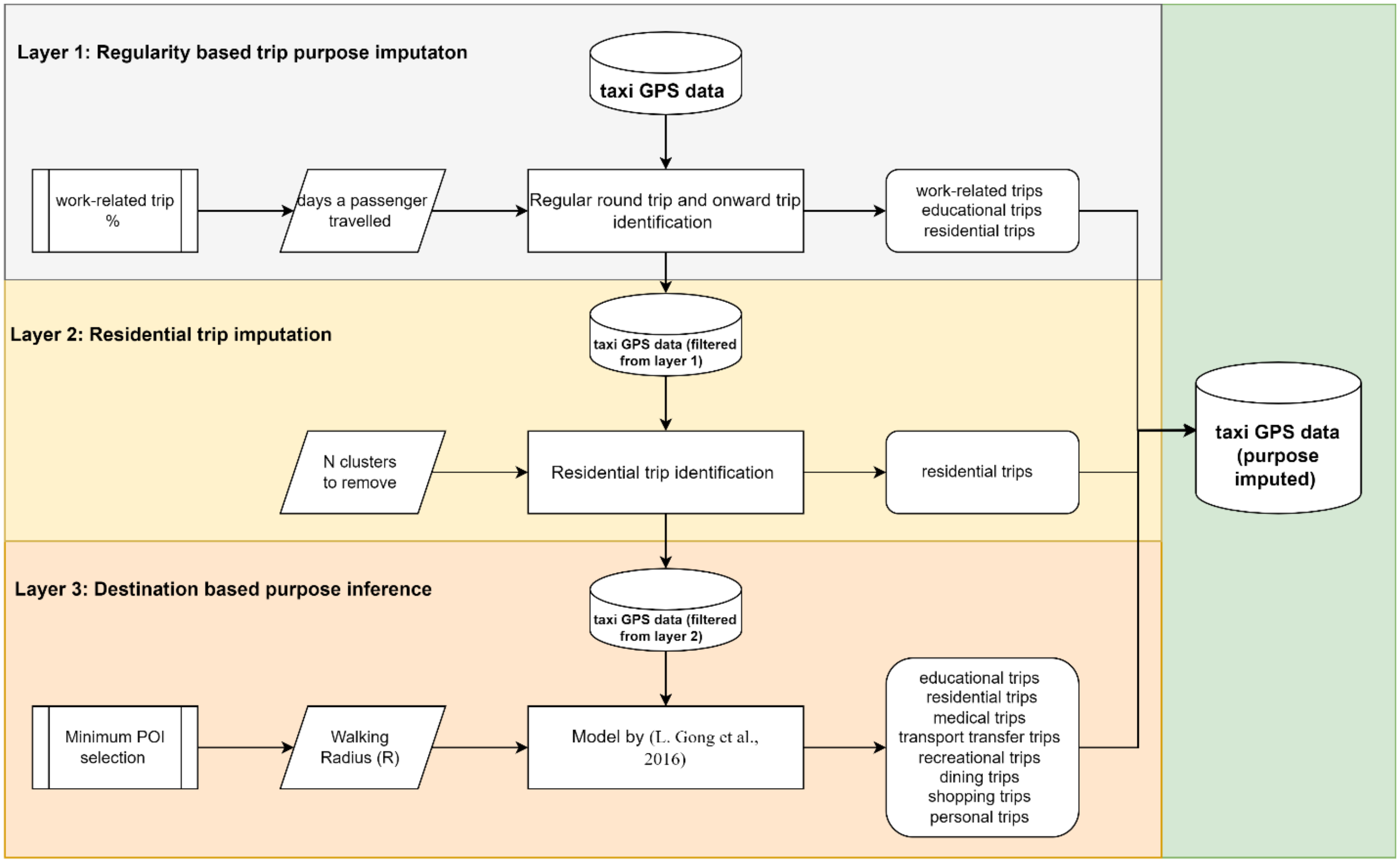

Figure 23: Designed three-layer framework for trip purpose inference

- Layer 2:
  - Label the trips attracted to residential places, to avoid inaccurate inference in absence of residential POIs (based on the constructed process in section 3.5).
  - In that process, since the step of distance-based clustering and selection of clusters impact to remove a higher number of trips, it was selected to optimize.
- Layer 3:
  - Followed by the first two layers, a filtered set of trips is inferred through the selected destination context-based trip purpose inference model.
  - Bayes-theorem-based model, which is based on the model of (Li Gong et al., 2016) was selected for this referred to the results shown in section 4.1.

3.8 Validating the Results

Validation of the designed three-layer framework for trip purpose inference is conducted with the assistance of household travel survey data. Achieved results were compared for the selected base model and the proposed three-layer framework in two ways as total activity proportion comparison and the DS division level comparison. Inter DS divisional trip purpose proportions for the selected three DS divisions of the study area from all the divisions within the Colombo District are considered for this requirement using the coefficient of the determination (adjusted R-Squared) value as the statistic. The adjusted R-Squared values tend to provide a more accurate perspective regarding the correlation between two variables (Miles, 2005). A complication that needed to be overcome in this context was the mismatch between the categorized trip purposes of this study and the travel survey. For this reason, only the shopping and educational purposes were selected to compare as similar practices had been taken in a few of the other trip purpose inference studies that use survey data for validation. (Li Gong et al., 2016).

3.9 Identifying the Travel Patterns

Three travel patterns were selected to identify through the designed purpose inference model as shown in below Table 8.

Table 8: Descriptions for travel patterns

Travel pattern	Description
Temporal regularities	Hourly distribution of the trip counts of each purpose for the selected week.
Spatial dynamics	Three-dimensional visualization through hexagons within the study area.
Trip length distribution	Distribution of trip lengths obtained from the OSRM for each purpose.

3.10 Limitations of the study.

It is worthy to mention that the applicability of the outcomes in this research is limit by two factors. The first is the land use context for which the proposed solutions only can be utilized in an area where a similar context exists in comparison to the selected area. This is caused by the attempt to fulfill identified one main research gap, that is trip purpose inference models have not been addressed in developing countries.

The second factor comes into account with the availability of the data. Four main data sources are used in this study namely taxi GPS data, POI data, road network data, and household travel survey data. In order to apply the proposed methods completely as in this study, the availability of these data is imperative. Hence, it is reliably suggested to consider these two main limitations in any attempt to apply this study to another context.

3.11 Chapter summary

- Taxi GPS data of a leading service provider in Sri Lanka and POI data from Google Places API are used as the two main datasets while route network data and household travel survey data are used as supplementary datasets.
- Three purpose inference models which focus on destination context are assessed comparatively. (1) Closest Assignment, (2) Gravity model, (3) Bayes theorem-based model.

- Collected POI data will be improved by two initiatives. (1) Increasing the number of informative POIs by trained SVM text classifier, (2) identifying the possible entering locations for selected POIs.
- Regularity-based trip purpose imputation will be conducted to identify regular round trips and onward trips.
- Three-layer trip purpose inference framework was designed for a streamlined purpose imputation process and will be validated at DS divisional level with the household travel survey.
- Three travel patterns (1) temporal regularities, (2) trip length distribution, and (3) spatial dynamics are evaluated for the purpose of imputed taxi trips.

4. RESULTS AND DISCUSSION

4.1 Selected Model Comparison Results

Table 9: Achieved purpose counts from the compared models

Purpose	Closest Assignment	Gravity Model	Bayes Model	Updated Bayes Model
Work	6,546	8,786	8,508	7,871
Personal	49,381	67,307	56,859	44,547
Shopping	17,170	14,095	24,556	23,682
Educational	2,973	829	1,809	2,626
Medical	9,556	6,915	7,563	9,878
Dining	10,007	5,691	3,619	13,084
Recreational	648	114	103	821
Mode Transfer	3,687	1,472	2,312	2,782
Multiple	5,381	110	20	58

Table 9 shows the total number of trips obtained for each purpose from the selected three-purpose inference models and the updated model that uses crowdsourced reviews for attractiveness. A notable fact is that a higher number of trips have been allocated for personal trips. It also aligns with having a higher number of POIs of personal category according to the followed matching. Gravity models' results have considerably been limited to this due to their structure in inferring the purposes with the number of POIs belonging to a purpose as shown in Equation 2. Important trip purposes such as educational and transport transfer is having the least values among the models for the gravity model due to this occurred limitation.

Not having the accessibility for a taxi or any other vehicle to reach the desired POI at all times is the practical difficulty that affects this. Hence, some of the studies argued for the accuracy of the closest assignment for the trip purpose inference problem (Furletti et al., 2013b). However, despite this argument, the obtained results show some plausible patterns than the Gravity model. Another complication of the closest assignment is having a substantial number of trips with multiple purposes which is shown as "Multiple" in Table 8. This has been largely caused by the use of a single parameter (distance from DOP to POI) for the purpose of inference. However, this issue has been reduced in the results of other models and the Bayes theorem-based models show the least values as per the use of many additional parameters. Due to the provided discrete values for the attractiveness value (A_i), Bayes model provided the least values for recreational and dining purposes. Thus, the importance of a more reliable measure for the attractiveness can be identified. This omission has reduced using the number of reviews for the attractiveness value in the updated Bayes model. This model has shown its capability to overcome the limitation of having a higher number of personal purposes as the depicted number of personal trips is the least among the models. Due to these observations among these compared models, the updated Bayes model can be selected as the more reliable model to be used for the selected study area.

4.2 Enhanced POI data Results

4.2.1 Predicted POI categories

Achieved classification results using the selected SVM classifier in predicting the category of a POI by its name are depicted in Figure 24.

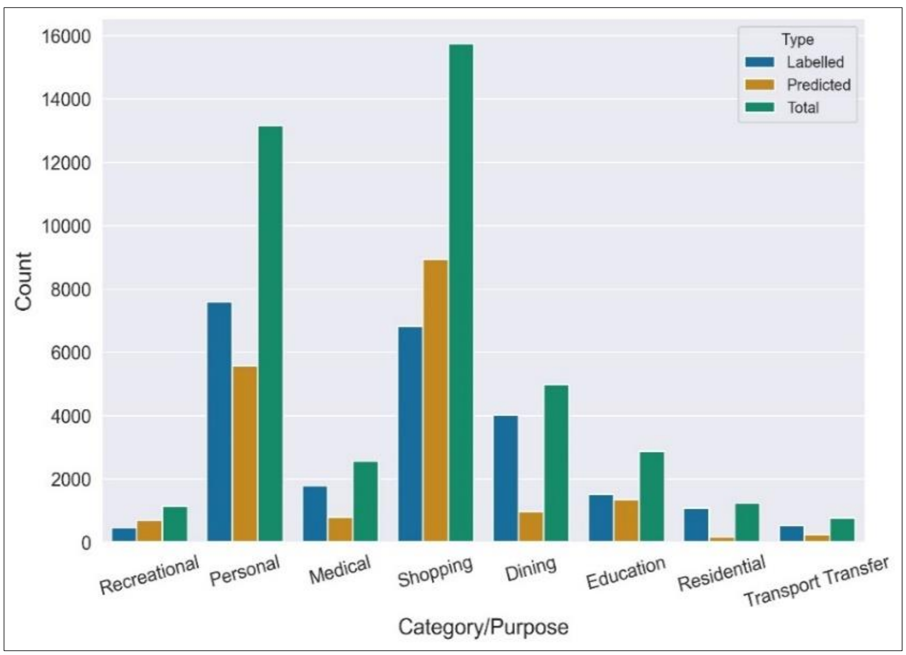

Figure 24: Total, Predicted vs labelled POIs for each purpose

The improvement in the number of categorized POIs can be seen from the total number of points for each category, which is obtained by combining the labeled and predicted instances. Figure 24 shows the distribution of labeled POIs and the distribution of total POIs (labelled + predicted) in the study area. A clear improvement in the POI density can be identified by comparing the two figures. It can be concluded that the level of

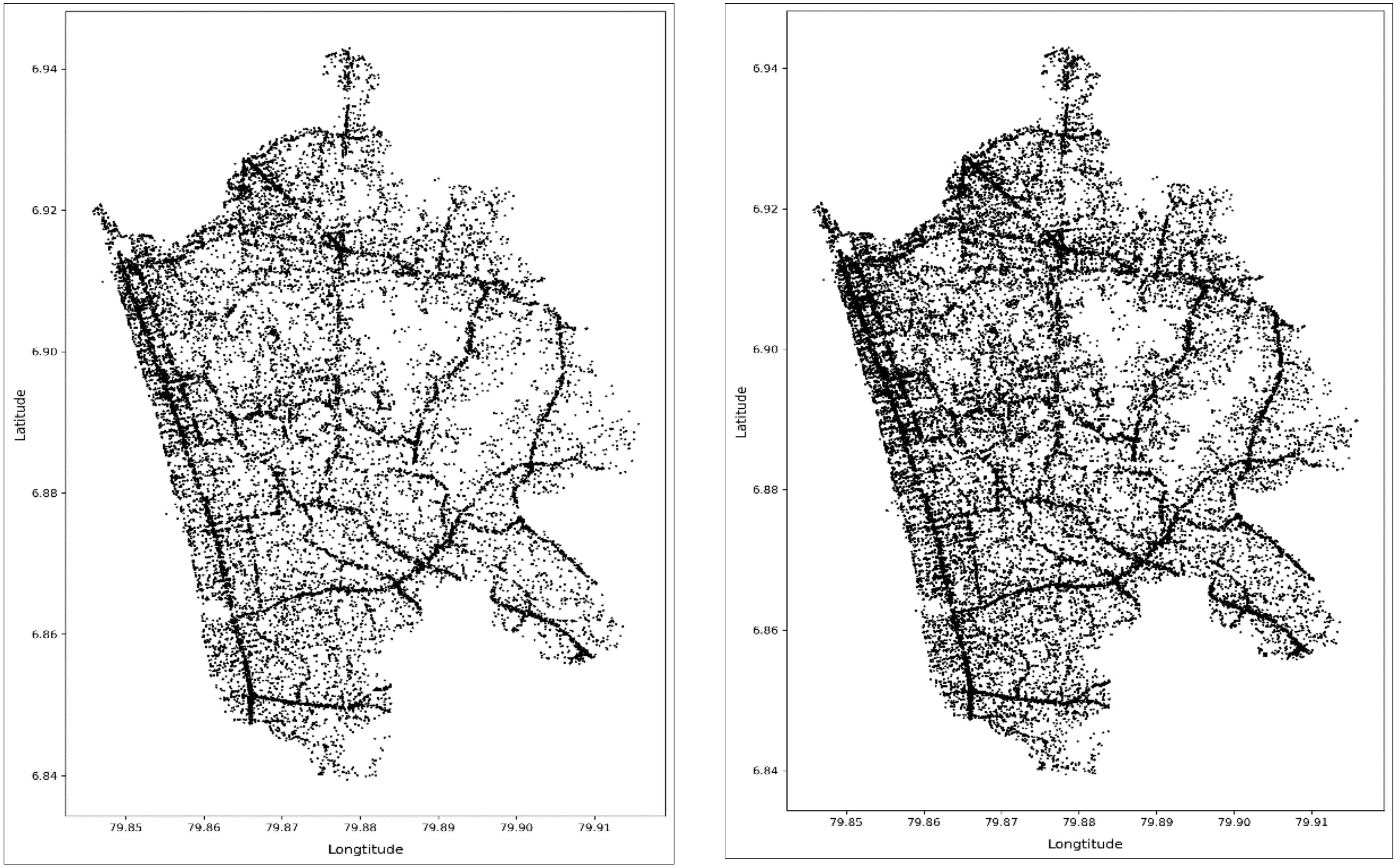

Figure 25: Spatial distribution difference between original and improved POI data
(left: original data, right: total data)

informative POIs have increased from 36.5% but with an accuracy around 76% due to the accuracy of the utilized SVM classifier.

4.2.2 Identified entering locations of POIs

As per the prerequisite of selecting the POIs that required entering locations, the use of k-means clustering for the attractiveness by "number of reviews" and selecting the clusters with higher attractiveness provided 4,332 POIs. The number of clusters was given as 15, aiming to obtain clusters with narrow data distributions, and the cluster selection was done by removing the clusters with a lower number of reviews. For this, the maximum number of reviews given per POI and the POI count of each cluster was used as the main parameters as shown in Table 10. Therefore, having maximum reviews of 28 and a substantial count of 38,805 POIs made the removal of only cluster no 1 in this scenario.

Following the same subsequently for the measured attribute "distance from the nearest road'" and selecting the clusters with higher distances reduced the selected number of POIs to 1,656 POIs. Table 11 depicts the achieved maximum "distance from the nearest road'" for each cluster with the POI count. The clusters with highlighted rows that have a lower distance were removed as per the defined threshold value of 50m. This scenario can be schematically visualized in Figure 26. Table 12 shows the count for each purpose from these selected POIs.

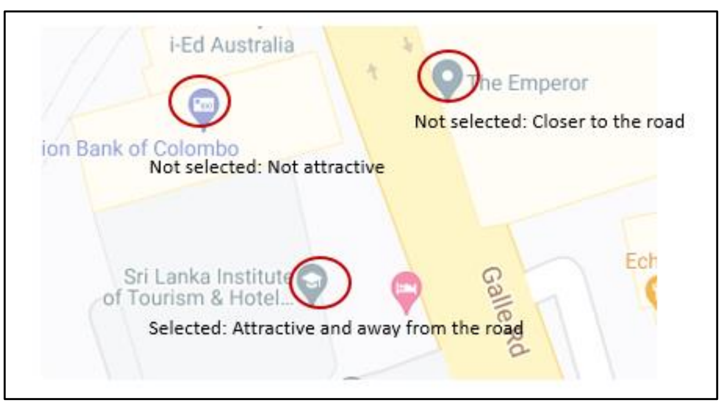

Figure 26: POI selection for entrance identification.

Table 11: POI and maximum review counts from attractiveness-based clustering

Cluster No	Maximum No of reviews per POI	POI count
1	28	38805
2	2059	57
3	11568	4
4	5502	12
5	678	217
6	18851	1
7	1004	140
8	7529	5
9	233	795
10	2867	26
11	103	2621
12	1471	82
13	3988	24
14	9088	2
15	421	347

Table 10: Maximum distance and POI counts from distance-based clustering

Cluster No	Maximum distance for a POI	POI count
1	28.25 m	668
2	86.5 m	124
3	60.5 m	287
4	143.25 m	45
5	8 m	324
6	52.25 m	476
7	223.5 m	13
8	72.25 m	189
9	21.75 m	641
10	35 m	600
11	122.25 m	55
12	177 m	21
13	50.5 m	353
14	15.5 m	443
15	103.25 m	93

Table 12: Purpose count for entrance location identified POIs

Purpose	**Selected POI count**
Shopping	501
Personal	482
Dining	267
Educational	148
Recreational	113
Medical	75
Residential	27
Transport Transfer	14

Taxi DOPs around the POIs were selected by providing the threshold as 200m, as a higher value than the given Walking Radius (R) of 100m is preferred to find the POIs that do not get selected by R. However, 400m was given for the POIs located very far from the roads as 200m may not be sufficient. The utilized DBSCAN algorithm to cluster DOPs was optimized by providing the parameters as 20m for the maximum sample distance and 10 for the minimum number of samples. The following Figure 27 shows an example selected entrance location for a POI through this designed process. Iterating this process for the selected set of POIs, 1,656 entrance locations were generated. Locations of these selected POIs were replaced by the identified entrance locations to use in the purpose inference. The distribution of the distance between the POIs and their entrance locations is shown in Figure 28. Although some percentage of entrance locations also has fallen for lower distances despite the POIs far from the roads were considered. Neglection of routes within the premises (e.g., university) can be the main possible reason for this observation.

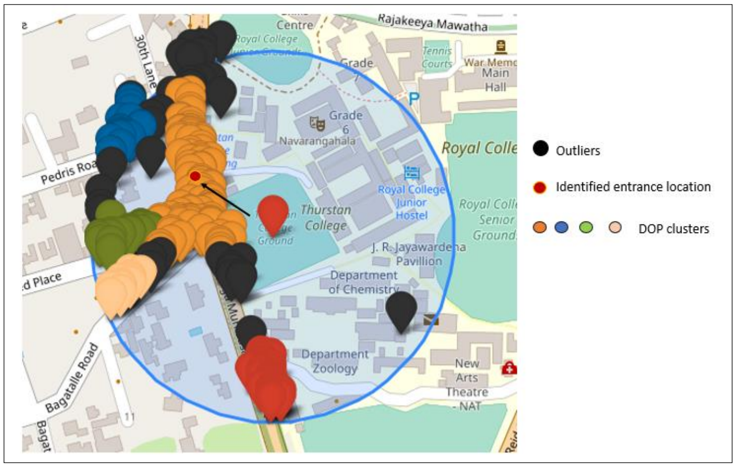

Figure 27: Sample entrance location identification

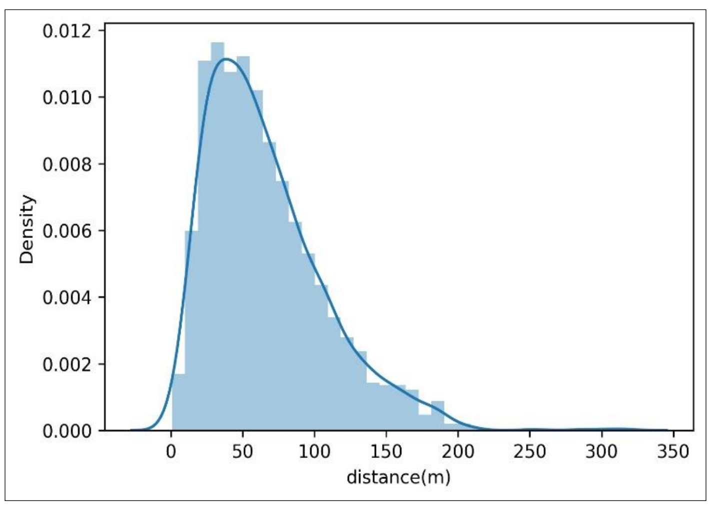

Figure 28: Distribution of distance between POIs and entrance locations

4.3 Purpose Imputation from the Three-Layer Framework

4.3.1 Regularity-based purpose imputation

As shown in Figure 23, the main parameter that defines the number of trips that will be assigned to purpose imputation through regularity is the minimum number of days a passenger should travel between identical locations to consider it as a set of regular trips. The daily percentage of regular work trips from household travel survey data for the considered study area was used to determine this parameter as suggested. Table 13 shows the achieved daily percentages of work trips from the developed regularity-based purpose imputation method for a selected set of minimum days.

Table 13: Identified work trip percentages from regular trips

Minimum days for regularity	% of work trips
2	7.2%
3	4.8%
4	2.2%
5	0.8%

However, the household travel survey data provided a value of 22% for work-related regular trips, which is somewhat higher than all the achieved values. Due to this reason, the minimum number of days with the highest regularity-based work trip percentage was selected, which is 2 days for this case. Despite that the expected percentage was not reached in this case, there is a huge possibility that the gap will be narrow when the period of trips is larger. For instance, when the dataset has a period of around one month, the number of regular passengers over a certain lower minimum number of trips should be considerably higher. As it eventually increases the number of regular trips, the percentage should be higher than the one achieved with the dataset of one week. Summary details of the imputed trip purposes for regular trips are shown in the below Table 14 and the daily counts are depicted in Figure 29.

Table 14: Passenger counts for purpose imputation from regularities

Detail	Count
Identified regular passengers	7,605
Passengers who didn't satisfy the purpose imputation criteria	5,675
Passengers with onward trips	1905
Passengers with round trips	58

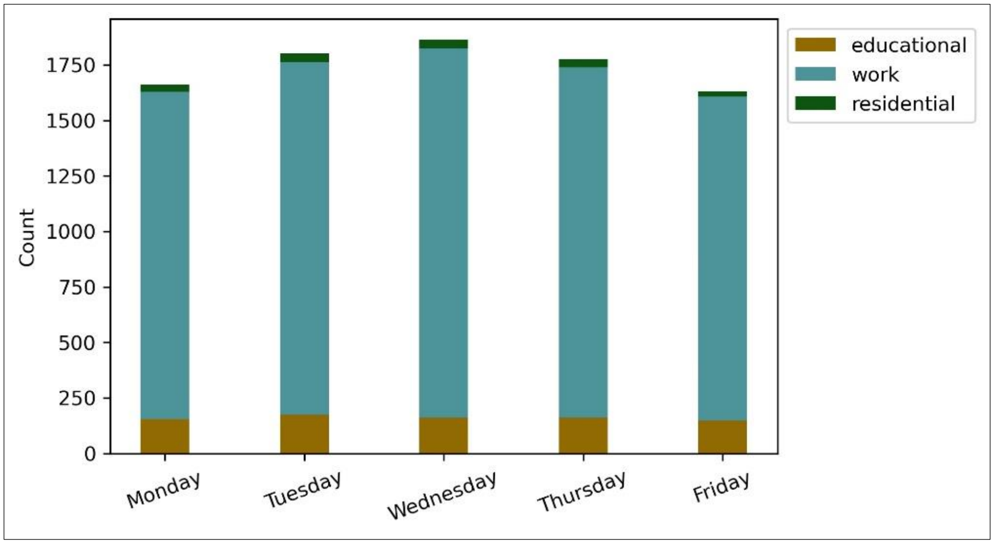

Figure 29: Daily distribution of imputed trip purposes from the layer 1

As per the assumption used to consider only the trips within the weekdays, the weekends are not subjected to these obtained results. Showing a small normal distribution between the days, the highest number of regular tips can be seen on Wednesday which has a value slightly upper than 1750. The identified educational trips are quite lower compared to the work trips, but it could be accepted as the children mostly used to travel by other means such as school vans. The number of residential trips imputed from regularity is less because only the round trips were used for the identification.

4.3.2 Residential trip imputation

The designed method to identify the trips attracted to residential places which are shown in Figure 18. was applied to the remaining set of taxi trips after the first layer. The method identified 13,816 trips as residential trips out of 98, 878. The following Table 15 shows the contraction of trips in applying the defined assumptions at different stages of the method.

Table 15: Trip count for stage-wise trip filtering in layer 2

Stage	Remained no of trips
Beginning	98, 878
Outlier selection	25, 862
Close distance cluster removal	19, 432
DOPs with near POI removal	13, 816

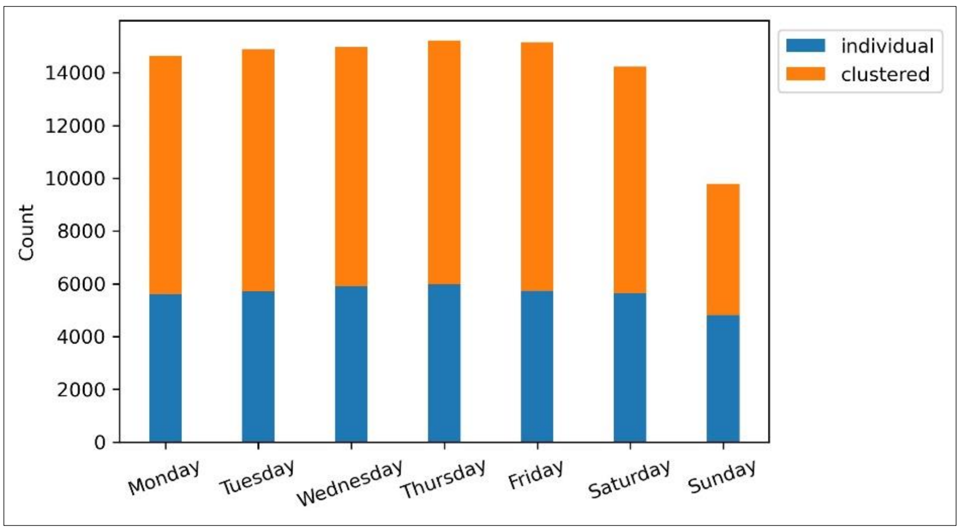

Figure 30: Daily distribution of clustered and individual trips in layer 2.

Clustering of the DOPs was done by using the minimum sample as 2 trips and eps(distance) as 50m for the DBSCAN algorithm. This provided the following daily distribution of grouped (clustered) DOPs and individual (outlier) DOPs as shown in Figure 30. The measured distribution of distances from the main roads was clustered into 15 clusters and the clusters lower than a minimum distance of 5m were removed as per the second assumption as shown in Table 16. Then for the rest of the remaining DOPs, a threshold of 15m was used to identify whether a categorized POI exists following the third assumption. This finally provided 13,816 taxi trips like the ones with the purpose of residential. Figure 31 shows the daily count of imputed residential trips from this second layer.

Table 16: Minimum distance from the main road and DOP counts from the distance-based clustering for layer 2

Cluster No	Minimum distance from the main road (m)	DOP count
1	20.40213	1,562
2	4.6789	6,430
3	87.0553	291
4	49.47306	766
5	127.1574	88
6	13.79548	2,757
7	8.544136	4,531
8	73.53362	450
9	6.882	6,235
10	38.57177	927
11	200.2985	4
12	103.3866	178
13	379.5019	1
14	61.11371	606
15	28.70671	1,036

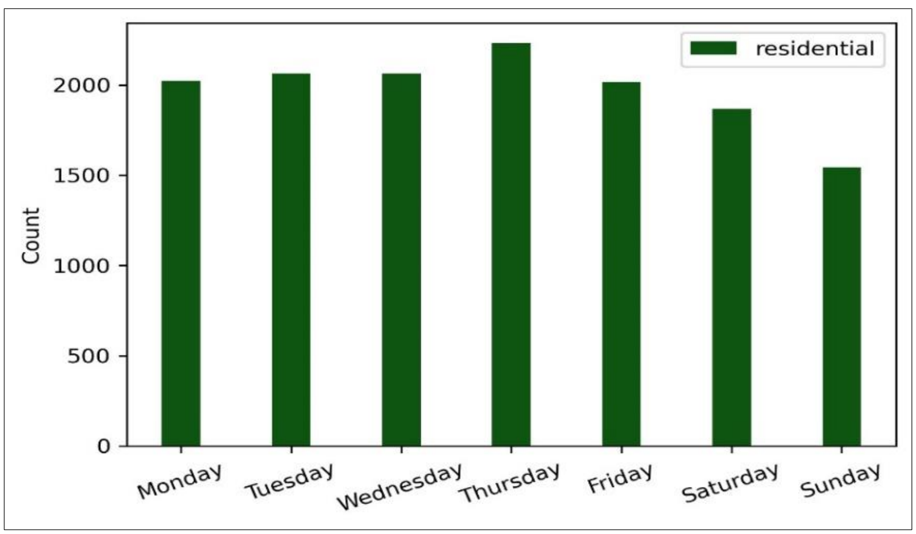

Figure 31: Daily distribution of identified residential trips from layer 2

4.3.3 Destination-based trip purpose inference

85,032 (79%) of filtered taxi trips after the first two layers of the three-layer framework were applied to the selected most suitable trip purpose inference model which is described under subchapter 4.1. Apart from the attractiveness factor, which was replaced by the number of reviews, the updated model from the proposed one by (Li Gong et al., 2016) the utilizes impact of time for the model. The used discrete values for the temporal impact that was determined by a small survey conducted for the study area are shown in Appendix 1. The following Figure 32 represents the achieved daily purpose count from the third layer.

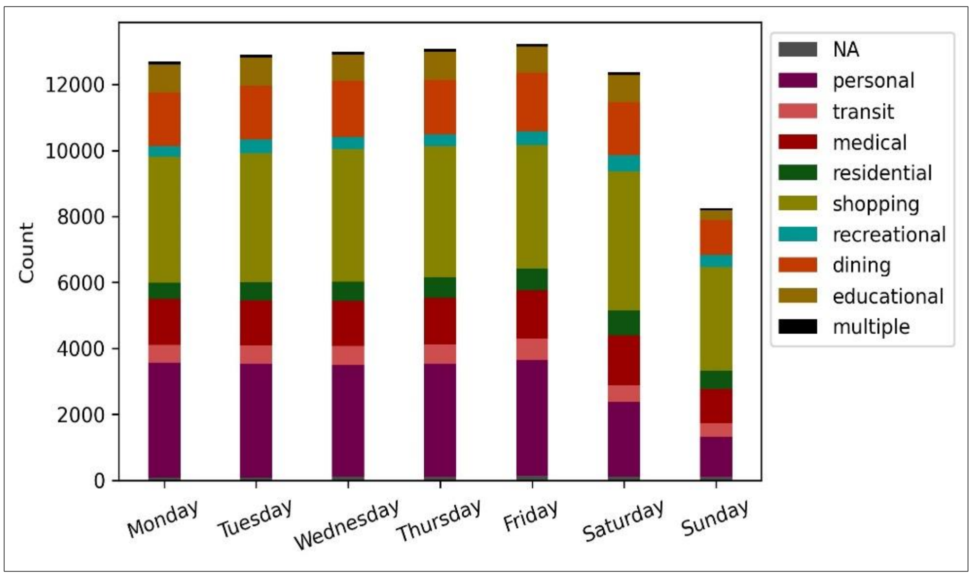

Figure 32: Daily distribution of trip purpose inference results from layer 3

A small number of trips have been assigned as "multiple" and "NA" due to the assignment of different POIs for the same DOP and the unavailability of the POIs at candidate POI selection. However, this number is almost negligible compared to the other proportions of trip purposes. The achieved count of daily trip purposes indicates that shopping is the highest followed by personal, dining, medical, and educational respectively as the main trip purposes. Moreover, recreational, transit, and residential have provided a relatively lower number of trips. A considerable difference between the proportions of purposes between different days cannot be observed apart from the total reduction of trips on Saturdays and Sundays due to the lower demand.

4.3.4. Complete results.

The obtained results from the proposed three-layer trip purpose inference framework are shown in the following Figure 33 as the daily count of trips.

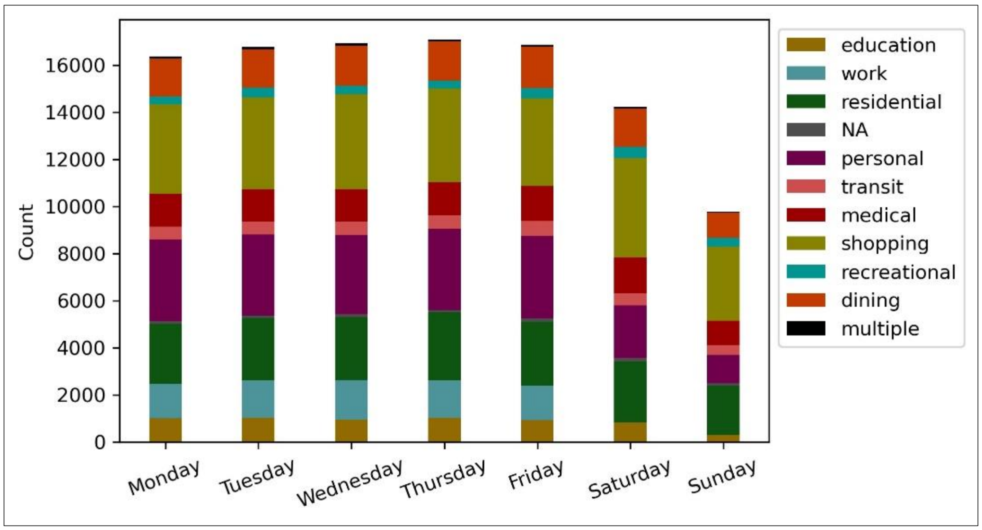

Figure 33: Daily distribution of trip purpose inference from the proposed three-layer framework

The addition of residential and work-related trips can be observed apparently in Figure 33 compared to Figure 32 in which the results of the first two layers are overlooked. An increment of around 4,000 trips for each day can be observed as a result of the complete three-layer framework rather than just using the existing mere bayes theorem-based model.

4.4 Model validation

It is a commonly accepted fact that the validation of these HG-based purpose inference models is rather difficult due to the nature of the study. However, in order to obtain a general overview of the accuracy of the model, the proportion of trip purposes was relatively compared with the household travel survey data. As the only purposes with non-overlapping content were the educational and shopping the comparison had to be limited. The following Table 17 depicts the comparative results only using the base model and the proposed model.

It can be identified that the gap between the trip proportions has been reduced to a considerable level by the proposed three-layer framework than the base model. Hence, these figures indicate the validity of the proposed model to a satisfactory extent.

Table 17: Trip proportion comparison from proposed model and base model

	Shopping (%)	Educational (%)
Household travel survey	25.2	14.5
Three-layer framework	24	12
Bayes theorem-based model	21	9

Despite the difficulty in validating the model precisely, another attempt was made based on the adjusted R-square values for the base and the proposed models. Since the household travel survey data provided the data at DS level, a division-wise comparison could be used for this requirement. The purpose overlapping issue between the two datasets limited this comparison also to educational and shopping purposes. The following Figure 34 depicts the achieved graphs and Table 18 shows the adjusted R-square values.

Table 18: adjusted- R-square value comparison between proposed and base model

	Educational	Shopping
Three-layer framework	0.524	0.518
Bayesian Model	0.343	0.431

Higher the adjusted R-square values higher the accuracy of the model. Hence, it is obvious that the proposed three-layer framework has provided a better result than the base model. However, it is notable fact that these adjusted R-square values are not satisfactory to a considerable level. This has been caused due to the limitations of the household travel survey data that were mentioned under section 3.2.4.

The distribution of the data points has been narrowed by the three-layer framework as indicated in Figure 34 and Figure 35, which is a solid indication of an achieved improvement by the model.

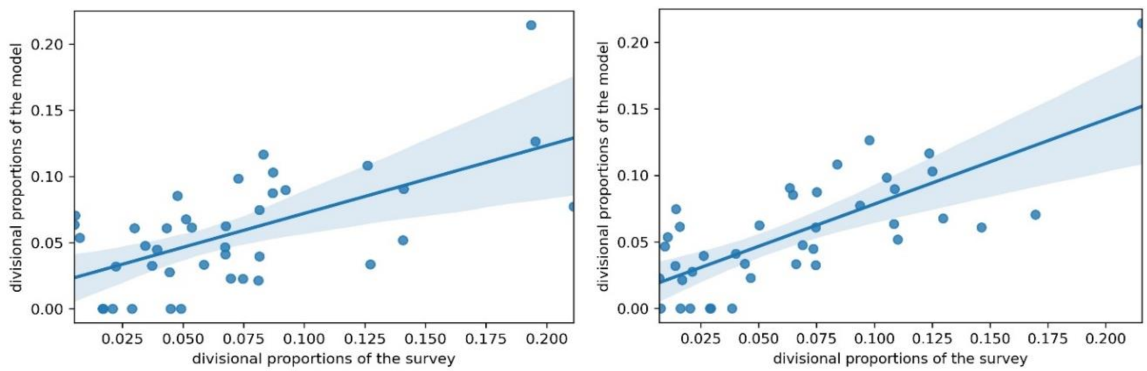

Figure 34: DS divsion level shopping trip proportion comparison between proposed (right) and base model (left)

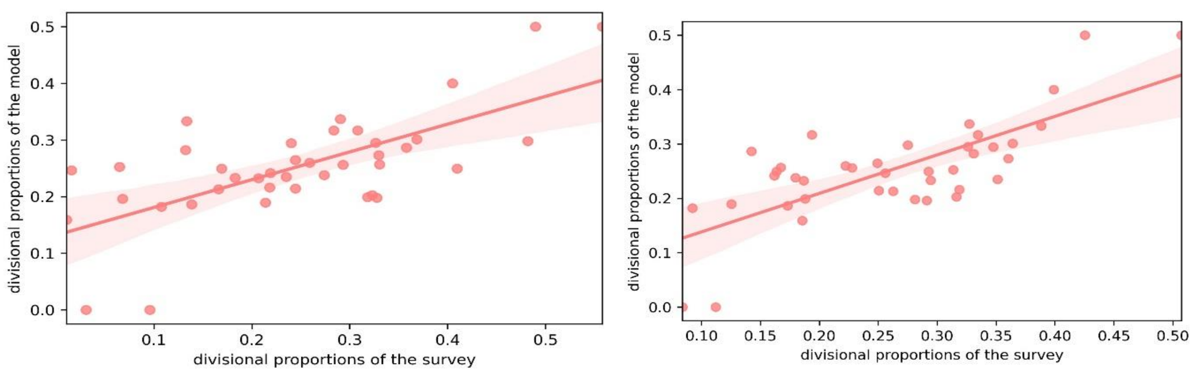

Figure 35: DS division level educational trip proportion comparison between proposed (right) and base model (left)

4.5 Identified Travel Patterns

4.5.1 Temporal distributions

As shown in Figure 36, the capability of the model to identify the work trips is limited to the weekdays. The peak in the morning around 9:00 a.m. can be identified regularly on each day. Also, there has been a slight increase in the afternoon around 2:00 p.m. which must have been caused because of the given time bin of the work trip identification in the first layer of the model.

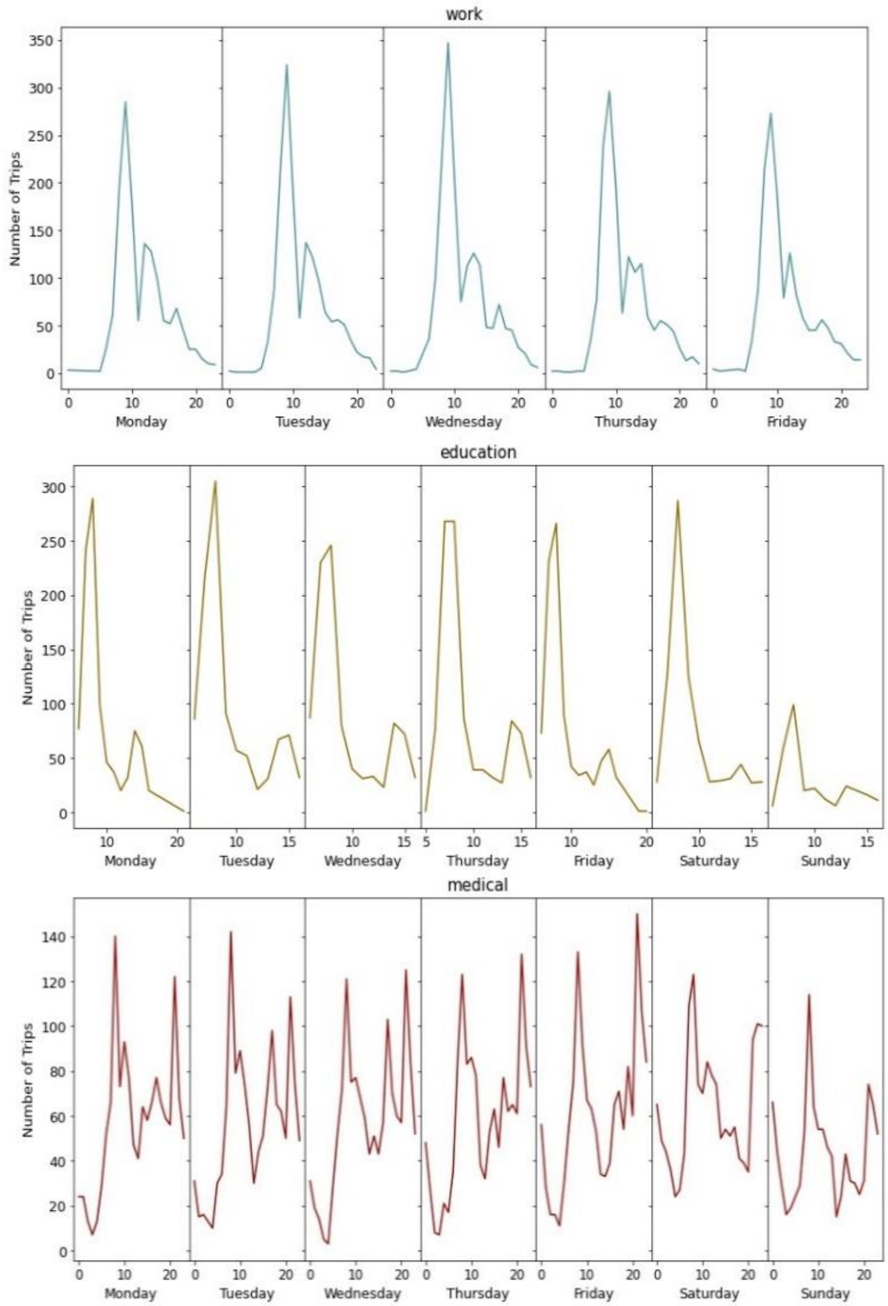

Figure 36: Temporal regularities of work, education and medical trips

The peak of the education trips has been recorded in the morning around 7:00 p.m. on the weekdays. However, during the weekends this has shifted slightly for a time around

9:00 a.m. This is mostly because on the weekdays the school start time is at 7:30 a.m. but on the weekends as the students tend to attend private educational institutes which often open later.

The temporal distribution of medical trips in Figure 36 indicates that people tend to visit for medical-related activities late at night as well compared to the morning.

Attendance for personal purposes such as service-oriented activities have a peak in the morning and also a solid distribution for the afternoon as well until the sudden drop in the evening as per the Figure 37. This must have been caused by the close of POIs related to personal activities in the evening around 5:00

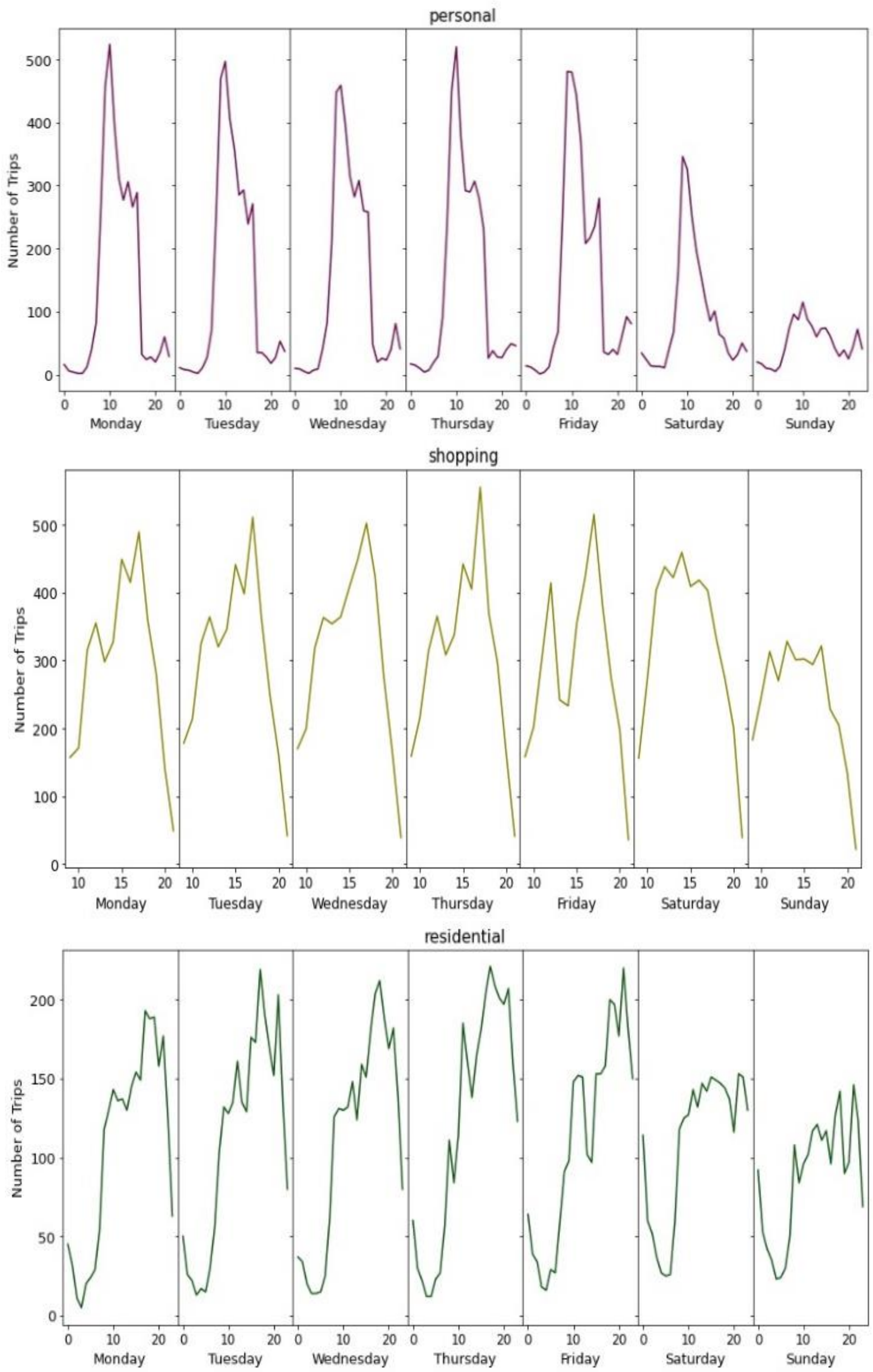

Figure 37: Temporal regularities of the personal, shopping and residential trips

p.m. In contrast to most of the other activities shopping, related trips have increased gradually from the morning and peaked in the afternoon around 6:00 p.m. Many people tend to be shopping after the work has finished which might be the reason behind that pattern. However, since the weekends' number of work trips is less, for the shopping that pattern cannot be observed.

Figure 37 indicates the attraction of trips to residential places are increasing at the evening and the peak has been recorded around 8:00 p.m. regularly. This is a common observation in any area where people tend to arrive home at night

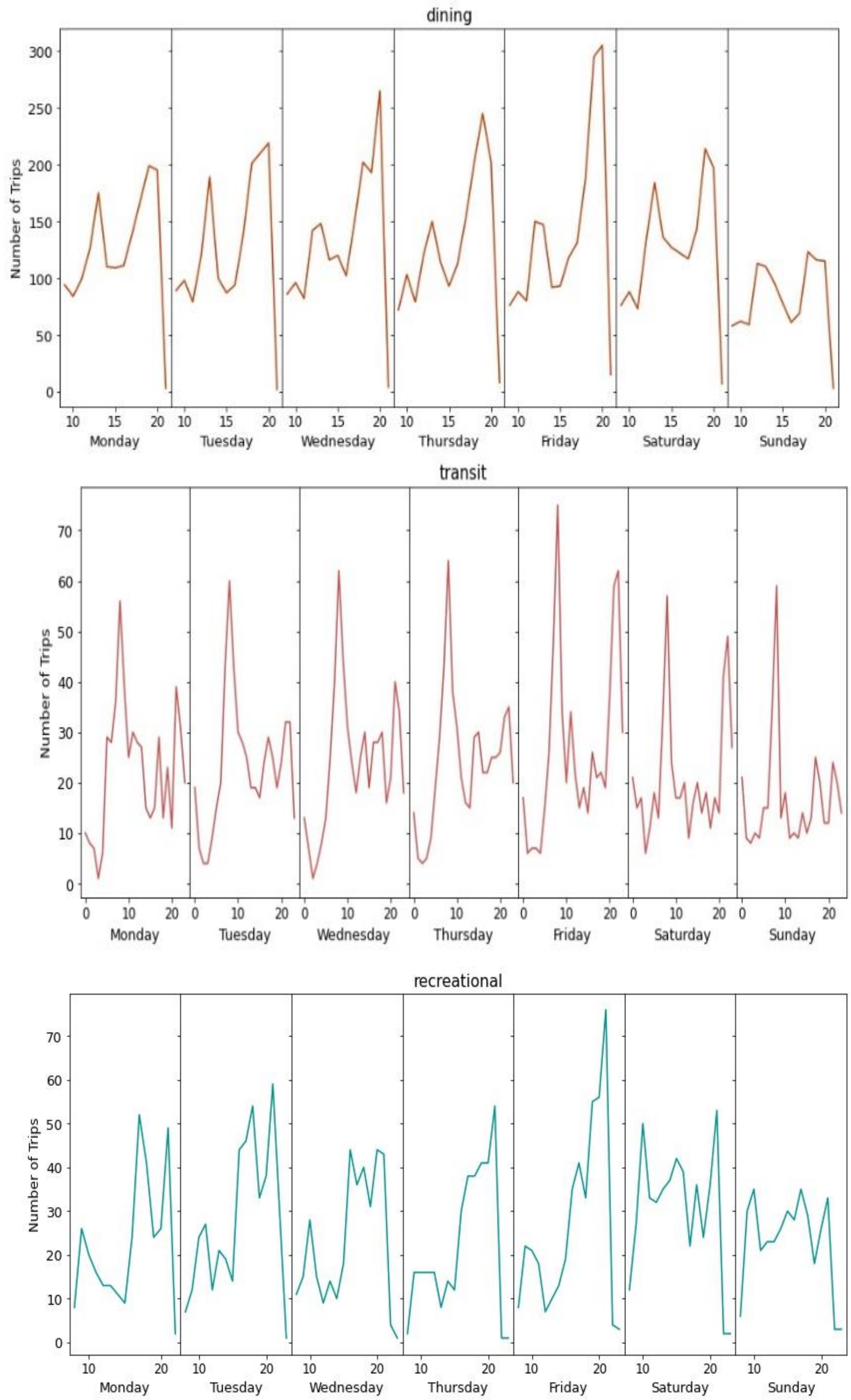

Figure 38: Temporal regularities of the dining, transit and recreational trips

after the work.

The dining activity which is concerned as a visit to purchase foods at any time of the day also gradually increases throughout the day and peaks at the nighttime as per Figure 38. The trips intended to transit or move to another mode of transportation tops at the morning around the period 7:00 – 9:00 a.m. as people mostly begin trips in the morning to complete longer distance trips. However, a solid distribution also can be seen throughout the day for this activity. The recreational activities show a somewhat doubtful pattern as the increase has been recorded at the evening time. However, this has been mitigated to a certain extent at the weekdays where a flat trend can be seen in Figure 38.

4.5.2 Trip length distribution

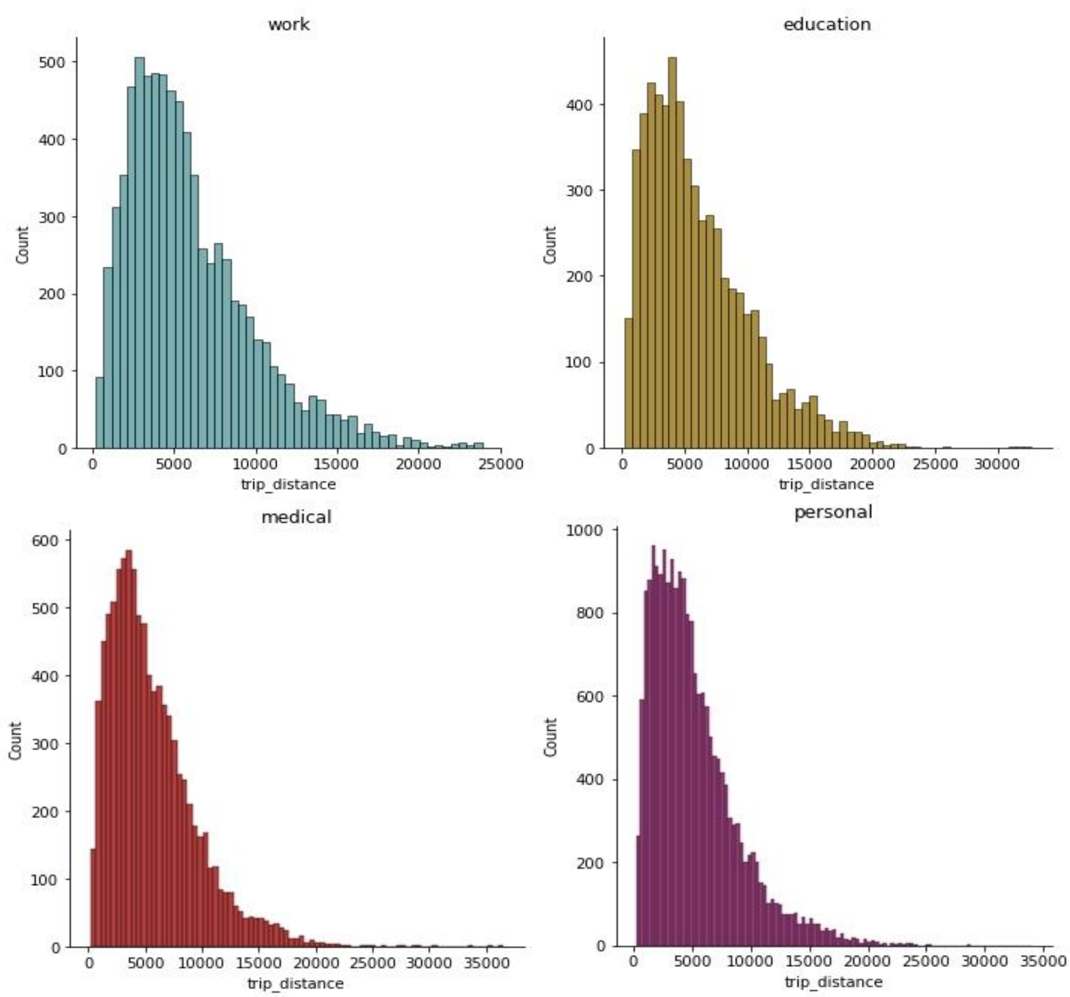

Figure 38: Trip length distributions for work, educational, medical and personal trips (in meters)

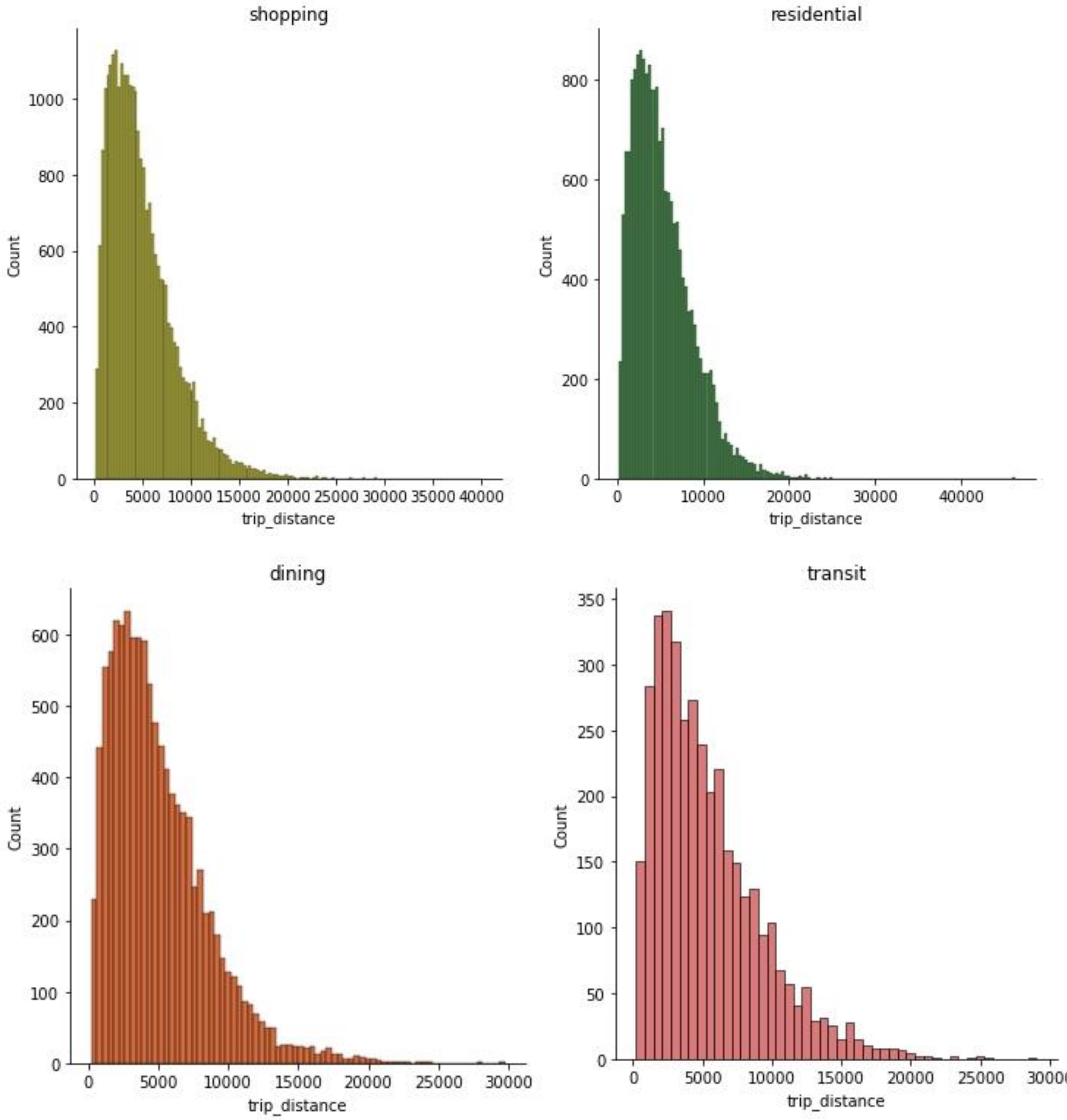

Figure 39: Trip length distributions for personal, residential, dining and transit trips (in meters)

Figures 38, 39, and 40 indicate the trip length distributions for all the purposes considered in this study. All the trip length distributions are skewed to the right side, which has been caused by the lesser number of longer trips. In terms of the counts, the peak of each purpose depends on the trip counts of those. It can be clearly identified from having higher peak values for purposes such as shopping and personal which are

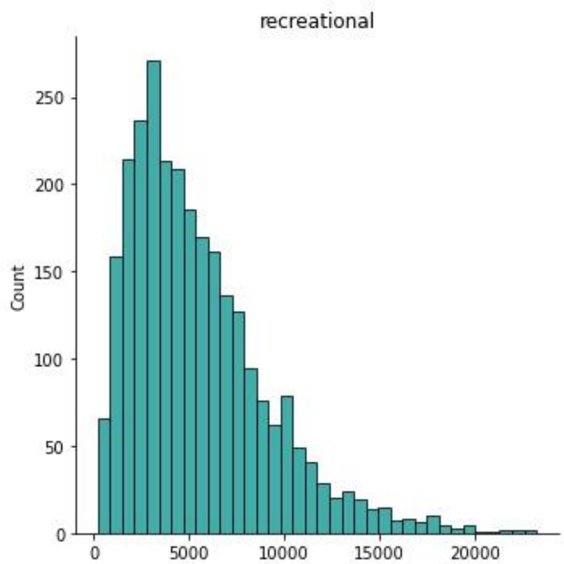

Figure 41: Trip length distribution for recreational trips (in meters)

activities with higher trip counts and having lower values for recreational and transit in contrast. The maximum value of the x-axis which indicates the trips with the highest distances varies with different purposes. For instance, some residential and shopping trips take more than 35Km while the maximum distance for recreational and work trips is around 25Km. However, it is worthy to mention that these maximum values could be the outliers in comparison to the total distribution. Hence, to obtain acceptable observations including descriptive statistics boxplots were created using the 1.5 Inter Quartile Range (IQR) rule for the trip lengths as shown in the below Figure 41. According to that, once the outliers are removed, the highest distance from all the purposes is recorded with educational trips which is a value just above 16Km.

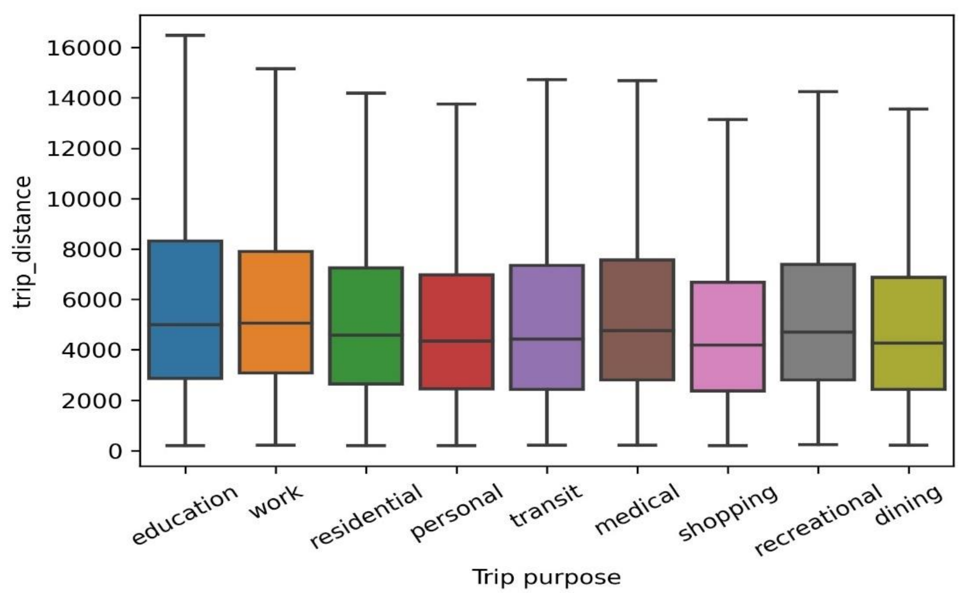

Figure 40: Box plots for the trip length distributions for each purpose

Table 18 highlights the mean trip length for each trip purpose received from the developed framework. In order to determine whether there is a significant difference between the distribution of trip lengths in activity-based trips (sample) and all trips (population), a one-sample z test was conducted which is depicted in Table 19. Furthermore, two-sample z tests were conducted for distributions between activity-based trip lengths for each purpose as shown in Table 20. (Test could be conducted even if the distribution is nonnormal as the sample size is larger and z-score standardization was used before the comparison, level of significance = 0.95)

Table 19: Mean trip lengths for purposes

	education	work	residential	personal	transit	medical	shopping	recreational	dining
Mean Trip Length	6032.56	5923.528	5321.480	5205.143	5379.769	5626.624	4948.501	5503.469	5053.077

Table 20: Z-test results for the samples vs population for trip lengths

	education	work	residential	personal	transit	medical	shopping	recreational	dining
Z-Value	15.48864	4.14E-54	14.93242	2.03E-50	0.868122	0.385328	-3.60643	0.00031	1.371006
P-Value	4.14E-54	2.03E-50	0.385328	0.00031	0.170373	3.07E-18	4.11E-54	0.00364	3.26E-12

A slight deviation between the mean trip lengths for different trip purposes can be identified according to the provided values in Table 19. The maximum mean length has been taken by the education trips and the minimum length has been taken by the shopping trips. However, these observed deviations may vary with the empirical concerns over the trip lengths of different activities. Hence, the conducted z values provide a sense of how these distributions vary. According to Table 20, only the educational and shopping provide acceptable z values (gather than 1.96) for having a different distribution than all the trips although considering the p-value (less than 0.05) it even limits to the educational trips. However, based on the z values and p values of Table 21, it can be concluded that these distributions significantly differ between each trip purpose.

Table 21: Z-test results for the samples vs samples for trip lengths

	education	work	residential	personal	transit	medical	shopping	recreational	dining
education	0	1.722418	12.98253	15.33869	8.547734	6.694219	20.63756	6.210611	16.59364
(P-value)	1	0.08499	1.54E-38	4.22E-53	1.26E-17	2.17E-11	1.26E-94	5.28E-10	7.75E-62
work	-1.72242	0	12.02782	14.61732	7.44254	5.262817	20.48848	5.107831	15.919
(P-value)	0.0849938	1	2.54E-33	2.18E-48	9.88E-14	1.42E-07	2.73E-93	3.26E-07	4.65E-57
residential	-12.9825	-12.0278	0	3.100369	-0.88575	-6.54239	10.51837	-2.39878	6.028223
(P-value)	1.54E-38	2.54E-33	1	0.0019328	0.37575	6.05E-11	7.11E-26	0.016449846	1.66E-09
personal	-15.3387	-14.6173	-3.10037	0	-2.68166	-9.22991	7.511738	-3.96336	3.495714
(P-value)	4.22E-53	2.18E-48	0.0019328	1	0.007326	2.71E-20	5.83E-14	7.39E-05	0.00047
transit	-8.54773	-7.44254	0.885755	2.681656	0	-3.48833	6.743648	-1.3348	4.708113
(P-value)	1.26E-17	9.88E-14	0.37574959	0.0073259	1	0.000486	1.54E-11	0.181942707	2.50E-06
medical	-6.69422	-5.26282	6.542391	9.229909	3.488327	0	15.41732	1.53548	11.12131
(P-value)	2.17E-11	1.42E-07	6.05E-11	2.71E-20	0.000486	1	1.25E-53	0.124665948	9.88E-29
shopping	-20.6376	-20.4885	-10.5184	-7.51174	-6.74365	-15.4173	0	-7.47295	-2.50582
(P-value)	1.26E-94	2.73E-93	7.11E-26	5.83E-14	1.54E-11	1.25E-53	1	7.84E-14	0.01222
recreational	-6.21061	-5.10783	2.398779	3.963364	1.334797	-1.53548	7.472953	0	5.701571
(P-value)	5.28E-10	3.26E-07	0.01644985	7.39E-05	0.181943	0.124666	7.84E-14	1	1.19E-08
dining	-16.5936	-15.9194	-6.02822	-3.49571	-4.70811	-11.1213	2.505821	-5.70157	0
(P-value)	7.75E-62	4.65E-57	1.66E-09	0.0004728	2.50E-06	9.88E-29	0.0122167	1.19E-08	1

4.5.2 Spatial dynamics

Figure 42: Base layer of the spatial dynamics' visualization

A three-dimension Hexagon plot was used to identify the spatial dynamics of taxi trips attracted to the selected study area. The width of a hexagon was given as 100m matching the provided radius in Bayesian purpose inference. The higher the height of a hexagon the more the trips attracted on a specific purpose for that area. The base layer of the visualization shows the distinct ds divisions as given in the Figure 42. Furthermore, this chapter reasons the achieved spatial dynamics of each purpose through the rest of the figures.

Figure 43: Spatial dynamics of work trips

As shown in Figure 43, Thimirigasyaya ds division has a higher density in attracted work trips, which is plausible as this division has a higher number of working places than the rest. However, the peak is shown in locations near this division from others which tended to indicate working places with higher number of employees.

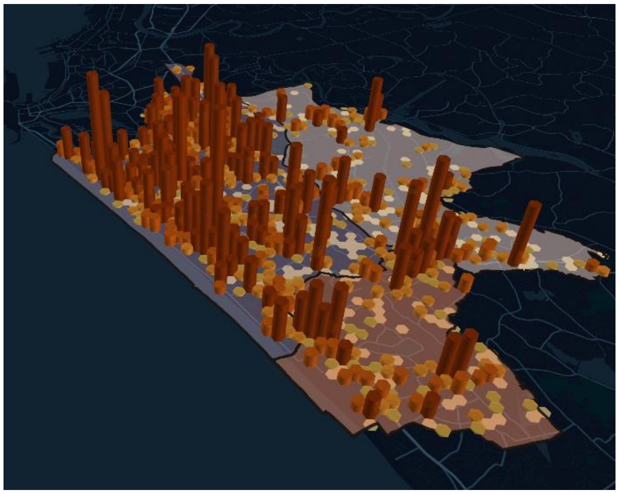

Figure 44: Spatial dynamics of educational trips

Figure 45: Spatial dynamics of medical trips

Figure 46: Spatial dynamics of personal trips

The distribution of educational activities is also mostly limited to the Thimbirigasyaya division according to Figure 44. The number of students in schools in this division is higher although most of the state-owned schools are distributed in all the divisions. Furthermore, most of the private educational institutes are located in this area. These two reasons may have impacted the received observation

The density of medical trips is quite low as discussed. However, a peak is shown in the Thimbirigasyaya division near the Kotte division where the hospitals such as 'Lanka Hospitals', and 'Army Hospitals are located which have a higher attraction of patients. Apart from those distributions with lower frequencies can be identified from the rest of the divisions.

Personal activities show a dense distribution in Figure 46 as one of the primary trip purposes. This has a higher density along the Galle corridor (left most corner of Thimbirigasyaya division) and also the region towards the Colombo

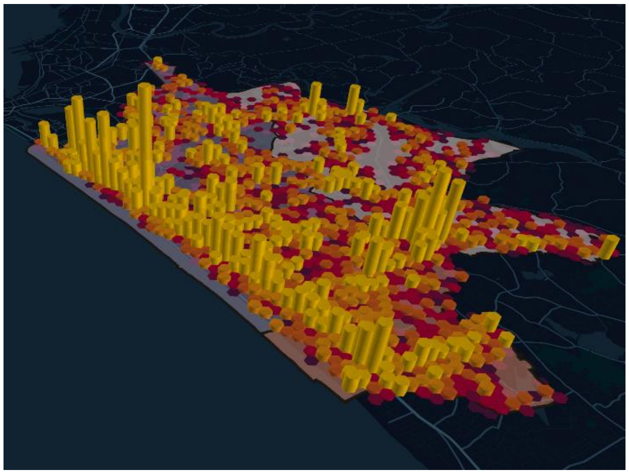

Figure 47: Spatial dynamics of shopping trips

Figure 49: Spatial dynamics of recreational trips

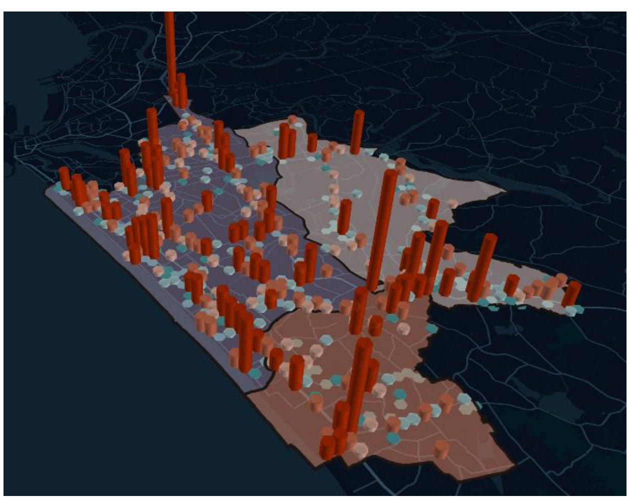

Figure 48: Spatial dynamics of transit trips

where most of the service-oriented Points of Interest (POIs) are located.

The distribution of shopping activities in Figure 47 also shows a generic pattern in attracting places with a higher number of POIs, which are located along the route network. However, a specific pattern cannot be identified for this purpose which has the highest number in POI distribution. The distribution of shopping POIs in most of the areas could be a possible reason for that.

The trip attraction of residential trips in Figure 48 shows a different pattern from the rest as expected. The higher densities are shown in the Dehiwala and Kotte ds divisions which lack attractive POIs than the Thimbirigasyaya. This is a reliable pattern as these two divisions consist with a higher number of residential areas as well.

The transit-oriented trips have been attracted to places near the main road network where the bus and train stations are located as per the Figure 49 . Hence, it shows a scattered distribution in the regions. Along the Galle Road, there are two lines can

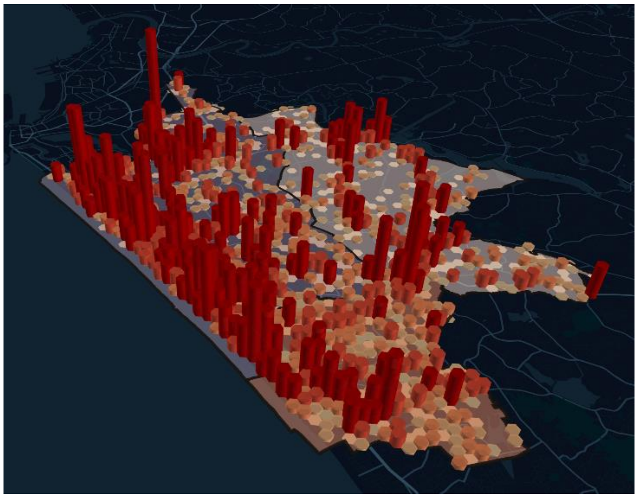

Figure 50: Spatial dynamics of dining trips

be seen one for train stations and the other for bus stations near the roads.

Dining activities in Figure 50 show a general spatial distribution. Mostly along with the areas in Galle Corridor and also the upper region in Thimbirigasyaya division. Most of the popular dining places are located along the main routes and a customer attraction factor can be the reason behind that.

Figure 51: Spatial dynamics of recreational trips

Recreational activities also show a scattered distribution as per the Figure 51. The peaks are shown around the middle of Thimbirigasyaya division where highly attractive recreational parks such as "Viharamahadevi" are located.

4.6 Chapter Summary

- Bayes-theorem-based model was selected as the most reliable one to use as a reference model for the considered land use context.
- The use of ML classification assisted to improve the level of informative POIs by 36.5% with an accuracy of 76% and the developed method to identify the entrance locations of selected POIs helped to mitigate a considerable number of possibly neglected POIs.
- The proposed three-layer framework for trip purpose inference, provided purpose inference results as 8.1% from the first layer considering the trip

regularities, 7.9% from the second layer focusing on trips attracted to residential places, and rest from the final layer.

- The validity of the proposed model could be identified from the increased archived R-square values for DS division level trip purpose proportion comparison.
- Attempt to identify the travel patterns through temporal regularities, trip lengths and spatial dynamics provided plausible patterns for the selected study area.
- It is recommended to further study the use of additional contexts such as trip origin neighborhood and trip time using UML techniques for the trip purpose inference.

5. CONCLUSION AND RECOMMENDATIONS

Automated data collection techniques have been welcomed for transportation analytics in overcoming the vast number of embedded deficiencies in widely used traditional techniques. Thus, AFC and GPS data have become the main two sources for demand-side data collection and GPS data are often used for identifying human mobility behaviors for their capability to capture people's movements from a fine granular level. Taxi is a transportation mode that facilities the travel needs of urban communities for decades and advent of the smart mobile devices caused to increase in its usage by creating a modern services such as on-demand taxis. On-demand taxi service providers collect the passenger trip data mainly through the provided origin and destination locations to their platform by the mobile user application, which uses mobile GPS technology. These large numbers of collected trip data are used by the service providers in revealing their knowledge regarding their customers. And this is also a highly informative medium for the ones who are interested in data science and transportation science fields. In this context, enrichment of these taxi data with the unavailable attributes is known as semantic enrichment, and the purpose of a trip is regarded as a difficult but highly valuable attribute that needs to be analyzed, and it created the trip purpose inference problem. Since developing countries often lag in advanced transportation analytics compared to developed countries, this problem hasn’t been addressed despite having different urban fabrics. To that end, this study focused on developing a suitable trip purpose inference framework using three DS divisions of Colombo District, Sri Lanka as the study area to extract informative activity-based travel patterns of taxi trips.

The accuracy and richness of the POI data are highly influential to the results of trip purpose inference. To that end as a key contribution to the existing knowledge, enhancement of POI data for the trip purpose inference is assessed in this study as one of the key contributions. The lack of informative POIs was addressed as one key issue of POI data, and the use of text classification to predict the category of a POI based on its name was evaluated. SVM providing an accuracy rate of 76.5%, was selected as the best ML classifier, and it was able to increase the number of informative POIs by around 35.5% from that. However, since this rate of accuracy largely impacts the results, it is encouraged to enhance this by deep learning architectures as a future study area. Another complication was the missing POIs to the given walking radius (R) around the destination context of a DOP. 1,651 such POIs were identified by using attractiveness and distance from roads to the POIs as main parameters, and the possible entering locations were identified using clustering for DOPs around those POIs in replacing the locations of those POIs in a way that is more suitable for the purpose inference.

Available existing trip purpose inference models were compared to select a base reference model in developing a suitable trip purpose inference model for the selected context. For this scenario, Rule-based: closest assignment, Probabilistic: gravity model and Bayes model were selected to compare due to the wide-scale usage. The closest assignment model tended to provide a higher number of multiple-purpose assignments for trips while the results of the gravity model tended to be biased for the available number of POIs for a purpose. Hence, the Bayes theorem-based model was selected to use as a reference model considering these two limitations. Furthermore, the use of an alternative data source to measure the attractiveness of POIs which is a key variable in the considered Bayes theorem-based model was tested utilizing the number of reviews for POIs. This helped to mitigate the impact of direct purpose inference of the original model from the use of discrete values and provided a smooth transition between different activities as a one-key contribution from this study.

The possible improvement can be made to obtain reliable trip purpose inference results using the selected model as the key reference considered. To this end, the inclusion of regular trips of passengers to impute the purposes was selected as an area to focus and it is included as the first layer of the proposed three-layer trip purpose inference framework. It is intuitively assumed that the passengers only take regular trips for work and educational purposes. However, since the time ranges that passengers travel for these two activities depend on the travel patterns within the area, using insights from a household travel survey dataset, time bins of 6:00 a.m. – 10: 00 a.m. and 12: 00 – 14: 00 p.m. was selected for the work trips and time bin of 6:00 – 9:00 a.m. was selected for educational trips. A regular trip could be either an onward trip (only for one direction) or a round trip (for both directions). The identification of regular trips was performed using the DBSCAN algorithm to cluster the origin and destinations and the hyperparameter eps (minimum number of samples) allows to determine the number of days a passenger should travel between identical origins and destinations. The percentage of regular work trips in the study region was used from the survey data to determine the optimal minimum number of days to consider a set of trips as regular in this layer. Although this optimality cannot be reached for the tested dataset which only consists of data for one week, it can be expected to be optimized for large datasets because the passengers captured to the proposed methodology increase by reducing the minimum days to consider. The first layer assisted to identify the purposes for 8.1% of the total taxi trips.

Another main concern was the inclusion of trips attracted to the residential places in the purpose inference since the POI datasets generally do not contain houses. Focusing on this concern the second layer of the framework was developed to identify the residential trips using three assumptions. The first is that residential places do not attract a higher number of trips per day which is captured from the outlier GPS records from the DBSCAN algorithm. The second is that residential places do not locate close to the main road network of the study area, and this is referenced by using k-means clustering for the measured distance of DOPs to main roads from OSRM and removing the clusters with shorter distances intuitively. Under the third assumption, if an attractive POI locates around the DOP that observation is also removed as there is a higher probability that a passenger may visit that place. Finally, this filtering procedure of the trips from the second layer provided 13, 816 (12.83%) trips as the residential trips for the utilized taxi trip dataset.

Filtered taxi trips data after the first and the second layer is intended to apply for the third layer in which the selected Bayes-theorem-based model is applied. 85,032 (79%) of trips for this considered dataset were applied for this third layer to infer trip purpose. However, for large-scale datasets, it can be expected that the proportions of the trips that are taken by the first two layers will increase leaving a somewhat lower proportion for the third layer with this increase in regular trips and residential trips. The validity of this proposed three-layer trip purpose inference framework is conducted in two methods using the original Bayes model as the base model. The proportions of work and education trips of taxi users are compared with the household travel survey data and the achieved reduction in activity proportions validated the model to a certain extent. Moreover, an R-square value comparison was conducted for the activity proportions at the ds division level. This highlighted that the proposed three-layer framework provided R-square values higher than the base model for the selected trip purposes stressing its validity.

The achieved trip purpose inference results from the proposed three-layer framework were utilized to identify the travel patterns of taxi users within the considered study area as temporal regularities, distribution in trip lengths, and spatial dynamics. Temporal regularities of the activity-based trips indicated the regular flow and peaks of trips for a week. Regular steady increment of educational and work trips in the morning, personal trips in the noon, shopping and residential trips in the afternoon, and dining trips in the evening are observed highlighting patterns. Apart from that, activities such as medical, transit, and recreational showed generic patterns with fluctuations. The lengths of the taxi trips were imputed utilizing the OSRM which provided the short distance between the PUP and DOP. The activity-based trip length distributions were right-skewed and showed only a slight deviation in between. However, the conducted two-sample z-tests for the mean values of activity-based trip lengths indicated that the difference is significant. Hence, it can be concluded that the educational trips have taken the highest lengths and the shopping has taken the least. Spatial dynamics for a different purpose were visualized three-dimensionally using a hexagon with a radius of 100m for the study area. The higher attraction of work and educational trips to the Thimbirigasyaya division and residential trips to the Dehiwala and Kotte divisions are highlighted spatial patterns at the macro level. Furthermore, increment in transit trips along the Galle corridor, shopping trips around areas with a higher number of POIs, and medical trips around popular hospitals also could be observed.

A number of future research directions can be identified from the limitations of this study. Although the proposed three-layer framework provided highly reliable results compared to the selected reference model, the tested dataset could not achieve its maximum potential. Hence, assessing the behavior of the model for a larger taxi trip dataset is required. There is a possibility to develop a machine learning-based weak supervision model to predict the trip purposes if a certain part of inferred trips can be identified as highly reliable using a developed criterion. This helps to perform the inference more conveniently and apply it to production-level tasks such as recommendation systems. It is more suitable to use the actual trip distances to assess the impact of trip length for different trip purposes rather than utilizing the short distance. This can be achieved by using the trajectory data of taxi trips. Although the destination context of the taxi DOPs was considered in the proposed model, there may be other factors that impact determining the purpose of a trip. Origin context of the taxi pick-up points and time of a trip are the two most important attributes that could make a substantial impact on the results. There is a possibility to impute these attributes through unsupervised machine learning techniques.

REFERENCE LIST

Bao, J., Xu, C., Liu, P., & Wang, W. (2017). Exploring Bikesharing Travel Patterns and Trip Purposes Using Smart Card Data and Online Point of Interests. *Networks and Spatial Economics*, *17*(4), 1231–1253. https://doi.org/10.1007/s11067-017-9366-x

Barrington-Leigh, C., & Millard-Ball, A. (2017). The world's user-generated road map is more than 80% complete. *PLoS ONE*, *12*(8), 1–20. https://doi.org/10.1371/journal.pone.0180698

Burbidge, S., & Goulias, K. (2009). Active travel behavior. *Transportation Letters*, *1*(2), 147–167. https://doi.org/10.3328/TL.2009.01.02.147-167

Chen, C., Jiao, S., Zhang, S., Liu, W., Feng, L., & Wang, Y. (2018). TripImputor: Real-Time imputing taxi trip purpose leveraging multi-sourced urban data. *IEEE Transactions on Intelligent Transportation Systems*, *19*(10), 3292–3304. https://doi.org/10.1109/TITS.2017.2771231

Chen, C., Liao, C., Xie, X., Wang, Y., & Zhao, J. (2019). Trip2Vec: a deep embedding approach for clustering and profiling taxi trip purposes. *Personal and Ubiquitous Computing*, *23*(1), 53–66. https://doi.org/10.1007/s00779-018-1175-9

Cui, Y., Meng, C., He, Q., & Gao, J. (2018). Forecasting current and next trip purpose with social media data and Google Places. *Transportation Research Part C: Emerging Technologies*, *97*(September), 159–174. https://doi.org/10.1016/j.trc.2018.10.017

De Jong, R., & Mensonides, W. (2003). Wearable Gps Device As a Data Collection Method for Travel Research. *Institute of Transport Studies Working Paper*, *ITS-WP-03-02*, 23 p.

Ermagun, A., Y. Fan, \, Wolfson, J., Adomavicius, G., & Das, K. (2017). Real-time trip purpose prediction using online location-based search and discovery services. *TRC*, *77*, 96–112. https://doi.org/10.1016/j.trc.2017.01.020

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 226–231.

Furletti, B., Cintia, P., Renso, C., & Spinsanti, L. (2013a). Inferring human activities from GPS tracks. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2505821.2505830

Furletti, B., Cintia, P., Renso, C., & Spinsanti, L. (2013b). Inferring human activities from GPS tracks. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August*. https://doi.org/10.1145/2505821.2505830

Gao, Q. ;, Molloy, J. ;, Axhausen, K. W., Gao, Q., & Molloy, J. (2021). *Trip purpose imputation using GPS trajectories with machine learning ETH Library Trip Purpose Imputation using GPS Trajectories with Machine Learning*. https://doi.org/10.3929/ethz-b-000505634

Goel, R. (2018). Distance-decay functions of travel to work trips in India. *Data in Brief*, *21*, 50–58. https://doi.org/https://doi.org/10.1016/j.dib.2018.09.096

Gong, Lei, Morikawa, T., Yamamoto, T., & Sato, H. (2014). Deriving Personal Trip Data from GPS Data: A Literature Review on the Existing Methodologies. *Procedia - Social and Behavioral Sciences*, *138*(0), 557–565. https://doi.org/10.1016/j.sbspro.2014.07.239

Gong, Li, Liu, X., Wu, L., & Liu, Y. (2016). Inferring trip purposes and uncovering travel patterns from taxi trajectory data. *Cartography and Geographic Information Science*, *43*(2), 103–114. https://doi.org/10.1080/15230406.2015.1014424

Gong, S., Cartlidge, J., Bai, R., Yue, Y., Li, Q., & Qiu, G. (2020). Extracting activity patterns from taxi trajectory data: a two-layer framework using spatio-temporal clustering, Bayesian probability and Monte Carlo simulation. *International Journal of Geographical Information Science*, *34*(6), 1210–1234. https://doi.org/10.1080/13658816.2019.1641715

González, M. C., Hidalgo, C. A., & Barabási, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, *453*(7196), 779–782. https://doi.org/10.1038/nature06958

Huang, L., Li, Q., & Yue, Y. (2010). Activity identification from GPS trajectories using spatial temporal POIs attractiveness. *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks, LBSN 2010 - Held in Conjunction with ACM SIGSPATIAL GIS 2010*, 27–30. https://doi.org/10.1145/1867699.1867704

IKONOMAKIS, M., KOTSIANTIS, S., & TAMPAKAS, V. (2005). Text Classification Using Machine Learning Techniques. *WSEAS TRANSACTIONS on COMPUTERS*, *4*(8), 966–974. https://doi.org/10.1007/978-981-33-4909-4_35

Lee, J., Yu, K., & Kim, J. (2021). Public bike trip purpose inference using point-of-interest data. *ISPRS International Journal of Geo-Information*, *10*(5). https://doi.org/10.3390/ijgi10050352

Lee, S., & Jang, K. (2019). Regularity of vehicle trips in urban areas. *2019 IEEE Intelligent Transportation Systems Conference, ITSC 2019*, 2651–2658. https://doi.org/10.1109/ITSC.2019.8917025

Li, A., Huang, Y., & Axhausen, K. W. (2020). An approach to imputing destination activities for inclusion in measures of bicycle accessibility. *Journal of Transport Geography*, *82*(January 2019), 102566. https://doi.org/10.1016/j.jtrangeo.2019.102566

Luo, X., Cottam, A., Wu, Y.-J., & Jiang, Y. (2021). Hybrid-Data Approach for

Estimating Trip Purposes. *Transportation Research Record: Journal of the Transportation Research Board*, *2675*(11), 545–553. https://doi.org/10.1177/03611981211018474

Meng, C., Cui, Y., He, Q., Su, L., & Gao, J. (2017). Travel purpose inference with GPS trajectories, POIs, and geo-tagged social media data. *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017*, *2018-Janua*, 1319–1324. https://doi.org/10.1109/BigData.2017.8258062

Merry, K., & Bettinger, P. (2019). Smartphone GPS accuracy study in an urban environment. *PLoS ONE*, *14*(7), 1–19. https://doi.org/10.1371/journal.pone.0219890

Miles, J. (2005). R-Squared, Adjusted R-Squared. In *Encyclopedia of Statistics in Behavioral Science*. John Wiley & Sons, Ltd. https://doi.org/https://doi.org/10.1002/0470013192.bsa526

Moiseeva, A., & Timmermans, H. (2010). Imputing relevant information from multi-day GPS tracers for retail planning and management using data fusion and context-sensitive learning. *Journal of Retailing and Consumer Services*, *17*(3), 189–199. https://doi.org/10.1016/j.jretconser.2010.03.011

Nguyen, M. H., Armoogum, J., Madre, J. L., & Garcia, C. (2020). Reviewing trip purpose imputation in GPS-based travel surveys. *Journal of Traffic and Transportation Engineering (English Edition)*, *7*(4), 395–412. https://doi.org/10.1016/j.jtte.2020.05.004

Pouke, M., Goncalves, J., Ferreira, D., & Kostakos, V. (2016). Practical simulation of virtual crowds using points of interest. *Computers, Environment and Urban Systems*, *57*, 118–129. https://doi.org/https://doi.org/10.1016/j.compenvurbsys.2016.02.004

Sari Aslam, N., Cheng, T., & Cheshire, J. (2019). A high-precision heuristic model to detect home and work locations from smart card data. *Geo-Spatial Information Science*, *22*(1), 1–11. https://doi.org/10.1080/10095020.2018.1545884

Shen, L., & Stopher, P. (2013). A process for trip purpose imputation from Global Positioning System data. *Transportation Research Part C: Emerging Technologies*, *36*, 261–267. https://doi.org/10.1016/j.trc.2013.09.004

Soares, E. F. D. S., Revoredo, K., Baiao, F., Quintella, C. A. D. M. S., & Campos, C. A. V. (2019). A Combined Solution for Real-Time Travel Mode Detection and Trip Purpose Prediction. *IEEE Transactions on Intelligent Transportation Systems*, *20*(12), 4655–4664. https://doi.org/10.1109/TITS.2019.2905601

Spaccapietra, S., Parent, C., Damiani, M. L., de Macedo, J. A., Porto, F., & Vangenot, C. (2008). A Conceptual View on Trajectories. *Data Knowl. Eng.*, *65*(1), 126–146. https://doi.org/10.1016/j.datak.2007.10.008

Tsui, S. Y. A., & Shalaby, A. S. (2006). Enhanced System for Link and Mode Identification for Personal Travel Surveys Based on Global Positioning Systems.

Transportation Research Record, *1972*(1), 38–45. https://doi.org/10.1177/0361198106197200105

Wang, P., Liu, G., Fu, Y., Zhou, Y., & Li, J. (2017). Spotting trip purposes from taxi trajectories: A general probabilistic model. *ACM Transactions on Intelligent Systems and Technology*, *9*(3). https://doi.org/10.1145/3078849

Wang, S., Li, L., Ma, W., & Chen, X. (2019). Trajectory analysis for on-demand services: A survey focusing on spatial-temporal demand and supply patterns. *Transportation Research Part C: Emerging Technologies*, *108*(September), 74–99. https://doi.org/10.1016/j.trc.2019.09.007

Wang, Y., Correia, G. H. A., de Romph, E., & Timmermans, H. J. P. (2017). Using metro smart card data to model location choice of after-work activities : An application to Shanghai. *Journal of Transport Geography*, *63*, 40–47.

Wolf, J. (2006). Applications of New Technologies in Travel Surveys. *Travel Survey Methods,* 531–544. https://doi.org/10.1108/9780080464015-029

Xie, K., Deng, K., & Zhou, X. (2009). From trajectories to activities: A spatio-temporal join approach. *GIS: Proceedings of the ACM International Symposium on Advances in Geographic Information Systems*, 25–32. https://doi.org/10.1145/1629890.1629897

Xing, Y., Wang, K., & Lu, J. J. (2020). Exploring travel patterns and trip purposes of dockless bike-sharing by analyzing massive bike-sharing data in Shanghai, China. *Journal of Transport Geography*, *87*(September 2019), 102787. https://doi.org/10.1016/j.jtrangeo.2020.102787

Yu, C., Li, H., Xu, X., Liu, J., Miao, J., Wang, Y., & Sun, Q. (2021). Data-Driven Approach for Passenger Mobility Pattern Recognition Using Spatiotemporal Embedding. *Journal of Advanced Transportation*, *2021*, 13–20. https://doi.org/10.1155/2021/5574093

Yue, Y., Wang, H., Hu, B., Li, Q., Li, Y., & Yeh, A. G. O. (2012). Exploratory calibration of a spatial interaction model using taxi GPS trajectories. *Computers, Environment and Urban Systems*, *36*(2), 140–153. https://doi.org/https://doi.org/10.1016/j.compenvurbsys.2011.09.002

Zheng, L., Xia, D., Zhao, X., Tan, L., Li, H., Chen, L., & Liu, W. (2018). Spatial–temporal travel pattern mining using massive taxi trajectory data. *Physica A: Statistical Mechanics and Its Applications*, *501*, 24–41. https://doi.org/10.1016/j.physa.2018.02.064

APPENDICES

Appendix 1: Questionnaire to measure the temporal impact for different trip

Impact of different times for taxi trip purposes within the Colombo District

This study attempts to find the impact of different time periods for the purposes of taxi trips within the Colombo District, Sri Lanka.

Please continue this survey if you have ever travelled by a taxi service within the Colombo District.

* Required

1. Have you ever travelled by a taxi service (i.e. PickMe, Uber) within the Colombo District?

Mark only one oval.

◯ Yes *Skip to question 2*

◯ No

Skip to question 2

For what kind of purposes do you travel by a taxi at different times?

In this questionnaire, a simple question is asked, that is, if you take a taxi trip at a particular time, what would be the most probable purpose of that trip?

Based on your experience of using taxi services, please tick the main trip purpose that you have fulfilled by taxi trips within each different time periods (Separately for Weekdays and Weekends).

Considered trip purposes are as follows.

1.) Education : (i.e. School)
2.) Work: (Your working place)
3.) Shopping: (i.e. Clothing store, Grocery store)
4.) Dining: (i.e. Restaurant)
5.) Recreational: (i.e. Movie theater)
6.) Personal: (i.e. Bank)
7.) Medical: (i.e. Hospital)
8.) Transport Transfer: (Use taxi as a mode to transfer for another transport mode (i.e. Train, Bus))

(note: if you do not travel in some time periods, please tick the last box "No")

purposes.

2. 1.) Please select the most probable activity that you may perform in each time frame. *

Check all that apply.

	Education	Work	Shopping	Dining	Recreational	Personal	Medical	Transport Transfer
00:00 - 02:00 a.m.	☐	☐	☐	☐	☐	☐	☐	☐
02:00 - 04:00 a.m.	☐	☐	☐	☐	☐	☐	☐	☐
04:00 - 06:00 a.m.	☐	☐	☐	☐	☐	☐	☐	☐
p.m.								
10:00 - 12:00 p.m.	☐	☐	☐	☐	☐	☐	☐	☐

Appendix 2: Determined discrete values to use in including the temporal impact to the bayes-theorem based model referring to the achieved results of the survey.

Time	Day	shopping	education	medical	recreational	personal	transit	dining	work	home
0:00:00	week	1	1	3	1	1	3	1	1	3
1:00:00	week	1	1	3	1	1	3	1	1	3
2:00:00	week	1	1	3	1	1	3	1	1	3
3:00:00	week	1	1	3	1	1	3	1	1	3
4:00:00	week	1	1	3	1	1	3	1	1	3
5:00:00	week	1	2	3	1	1	3	1	2	2
6:00:00	week	1	5	3	1	2	4	1	3	2
7:00:00	week	1	5	3	2	3	4	2	5	2
8:00:00	week	1	4	3	2	4	4	2	5	2
9:00:00	week	2	3	3	2	4	4	2	3	2
10:00:00	week	2	2	3	2	4	3	2	2	2
11:00:00	week	3	2	3	2	4	3	2	2	2
12:00:00	week	4	2	3	2	4	3	3	2	2
13:00:00	week	4	2	3	2	4	3	4	2	2
14:00:00	week	4	4	3	2	4	3	3	2	2
15:00:00	week	5	4	3	2	4	3	3	2	3
16:00:00	week	5	2	3	4	4	4	3	2	3
17:00:00	week	5	2	3	4	1	4	3	1	3
18:00:00	week	4	1	3	4	1	4	4	1	4
19:00:00	week	3	1	3	4	1	4	4	1	4
20:00:00	week	2	1	3	4	1	3	4	1	4
21:00:00	week	2	1	3	3	1	3	2	1	4
22:00:00	week	1	1	3	3	2	3	1	1	3
23:00:00	week	1	1	3	2	2	3	1	1	3
0:00:00	sat	1	1	3	1	1	3	1	1	3
1:00:00	sat	1	1	3	1	1	3	1	1	3
2:00:00	sat	1	1	3	1	1	3	1	1	3
3:00:00	sat	1	1	3	1	1	3	1	1	3
4:00:00	sat	1	1	3	1	1	3	1	1	3
5:00:00	sat	1	2	3	1	1	3	1	2	2
6:00:00	sat	1	3	3	1	2	3	1	3	2
7:00:00	sat	1	3	3	2	2	3	2	5	2
8:00:00	sat	1	4	3	3	3	3	2	5	2
9:00:00	sat	2	4	3	3	4	4	2	3	2
10:00:00	sat	3	3	3	3	4	3	2	2	2
11:00:00	sat	4	2	3	3	3	3	2	2	2

12:00:00	sat	4	2	3	3	3	3	3	2	2
13:00:00	sat	4	2	3	3	3	3	4	2	2
14:00:00	sat	4	2	3	4	2	3	3	2	2
15:00:00	sat	5	2	3	4	2	3	3	2	3
16:00:00	sat	5	2	3	4	2	4	3	2	3
17:00:00	sat	5	2	3	4	2	4	3	1	3
18:00:00	sat	5	1	3	4	2	4	4	1	4
19:00:00	sat	4	1	3	4	1	4	4	1	4
20:00:00	sat	3	1	3	4	1	3	4	1	4
21:00:00	sat	2	1	3	3	1	3	3	1	4
22:00:00	sat	1	1	3	3	1	3	2	1	3
23:00:00	sat	1	1	3	2	1	3	1	1	3
0:00:00	sun	1	1	3	1	1	3	1	1	3
1:00:00	sun	1	1	3	1	1	3	1	1	3
2:00:00	sun	1	1	3	1	1	3	1	1	3
3:00:00	sun	1	1	3	1	1	3	1	1	3
4:00:00	sun	1	1	3	1	1	3	1	1	3
5:00:00	sun	1	1	3	1	1	3	1	2	2
6:00:00	sun	1	1	3	1	2	4	1	2	2
7:00:00	sun	1	2	3	2	2	4	2	3	2
8:00:00	sun	2	2	3	3	2	4	2	3	2
9:00:00	sun	3	2	3	3	2	4	2	2	2
10:00:00	sun	3	2	3	3	2	3	2	2	2
11:00:00	sun	4	2	3	3	2	3	2	2	2
12:00:00	sun	4	2	3	3	2	3	3	2	2
13:00:00	sun	5	2	3	4	2	3	4	2	2
14:00:00	sun	5	2	3	4	2	3	3	2	2
15:00:00	sun	5	2	3	4	2	3	3	2	3
16:00:00	sun	5	1	3	4	2	4	3	2	3
17:00:00	sun	5	1	3	4	2	4	3	1	3
18:00:00	sun	4	1	3	4	2	4	4	1	4
19:00:00	sun	4	1	3	4	2	4	4	1	4
20:00:00	sun	3	1	3	4	2	3	4	1	4
21:00:00	sun	2	1	3	3	2	3	2	1	4
22:00:00	sun	1	1	3	3	2	3	1	1	3
23:00:00	sun	1	1	3	2	2	3	1	1	3