Trip Purpose: Data, Methods and Applications

LUCAS FELIX and ANTONIO LOUREIRO*, Instituto de Ciência Exatas, Brazil

Additional Key Words and Phrases: Mobility Analytics, Machine Learning, Analytics, Mobility

ACM Reference Format:
Lucas Felix and Antonio Loureiro. 2023. Trip Purpose: Data, Methods and Applications. *J. ACM* 0, 0, Article 000 (December 2023), 24 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Human mobility can be impacted by different aspects such as geographical, temporal, and social. As stated in [65], human mobility is part of what it means to live in the 21st century. Currently, loads of data are made available by ubiquitous computing, GPS-integrated devices, Call Detail Records (CDR), and Social Networks (SN), allowing the verification of human mobility patterns in practice [58]. Estimations based on digital data sources can help make faster and more informed decisions [65], which makes mobility information even more precious. Nevertheless, characterizing, evaluating, and analyzing these scenarios was not an easy task, due to the scarcity of data accessible, and when accessible most data do not present any contextual information. Contextual mobility information plays a significant role in understanding human mobility. Even though, most of the paths made by the user may seem random a first sight [69], when evaluating it closer it is possible to notice that people have a clear pattern. With contextual information, it is possible to increase the knowledge about the paths, providing predictions, recommendations, and plans. One problem though is that mobility can be impacted by several factors. According to [65], three main factors can impact mobility, spatial, temporal, and the reason for mobility. While the first two have been deeply studied, the third one, which is also known as Trip Purpose (TP), has been understudied [61].

Trip purpose can be defined as the goal of travel made by the user, i.e., the reason why the user is moving. In general, the TP of a trip can vary according to the work, nevertheless, it is common to see 8 classification purposes: home, education, shopping, eating out, recreation, personal, work, and transportation [61]. These purposes were defined by the California Householding Survey [9].

As stated in [58], TP is very important to estimate travel demands [11], and understand travel behaviors [20, 75]. It can deeply impact scenarios of transportation systems and urban planning [11], especially when considering the daily trips made by people. When thinking about the advantages to the users, TP can be used to improve users' characterization, consequently improving the quality of Recommender Systems (RS) and users' satisfiability [18].

A Trajectory, or Trip-Chain, (Tr) is a sequence of spatial points l_i with timestamps t_i, $Tr : (l_1, t_1) \rightarrow (l_2, t_2) \rightarrow ... \rightarrow (l_n, t_n)$ where each point l is represented by a pair of GPS coordinates, i.e., longitude and latitude. Thus, we regard "trip" as the movement from one location to another, e.g., $Tr_i \rightarrow Tr_{i+1}$, and we refer to the GPS points in each $l_i\ inTr_i$ as trip end locations. Thus, we

Authors' address: Lucas Felix, lucas.felix@dcc.ufmg.br; Antonio Loureiro, loureiro@dcc.ufmg.br, Instituto de Ciência Exatas, Belo Horizonte, Minas Gerais, Brazil.

0004-5411/2023/12-ART000 $15.00
https://doi.org/XXXXXXX.XXXXXXX

define the trip purpose as the activity that a user performs at a trip end location. Thus, the task of predicting the TP consists of identifying the motive why a person is visiting such a place and can be modeled in two ways. First, given a single trip Tr_{i+1}, the aim is to identify the purpose of one single travel, from $Tr_i \rightarrow Tr_{i+1}$. Secondly, given a Trajectory Tr, the aim is to identify the general purpose of the trips, or, to identify the purpose of the following trip Tr_{i+1}, using the features from the last i visits. However, as stated in [58], "in most cases, we cannot know the activities performed in a location without users' input".

Nevertheless, inferring the TP is not an easy task [61]. Different from other mobility-related tasks, TP can be an ambiguous task, as people can have different purposes in the same place. This is given by the inherent randomness of individual activity and the variation of trajectories from day to day and from individual to individual [27, 88]. Thus, users with the same mobility traces can have different intentions. Illustrating that, a person can go to a restaurant for a work-related meeting, while another person can visit the same restaurant with leisure intention. Thus, the trip purpose can be determined and impacted by several factors such as transportation type (e.g. bus, car, on foot), temporal aspects (e.g. visiting hour, day of the week), climate (e.g. is sunny/raining), the distance (e.g. if a person is hungry and on foot, would this person go to a distance place?), events, traffic, previously made visits, company (e.g. family, friends, alone), and categories visited [58]. Nevertheless, one problem commonly found in mobility datasets is the lack of some of these characterizing features or even the labels for the trip purpose or the travel classification. Hence, to properly accomplish such a task, a large amount of data from external is needed [61], contextual data. However, with the amount of data made available, identifying the appropriate data source for the task can also be arduous, given these datasets' different limitations.

To overcome the problems due to the data quality, works in the literature focus on different techniques, from more simple approaches as inference rules [66] to more complex approaches using supervised and unsupervised Machine Learning (ML) [76]. Each technique is applied according to the work' main goal, and the data available. Illustrating that, if defining people's work and home is enough, then, a simple inference rule can work well. In scenarios where the data is labeled, i.e., has the travel purpose, then a supervised ML can be applied. However, if no label is available, unsupervised ML is usually applied. Nevertheless, most of the proposed techniques are tailored to their data sources and applications, which makes it hard to adapt to other scenarios. The botton line is that understanding the goal of the application, the data available, and the techniques that could be applied given these two factors can be hard.

Given these facts, our aim with this survey is to organize the literature on trip purposes. Given the many applications that could benefit from the TP information, we give an overview of the literature, enabling researchers and practitioners to determine the most suited data and method for a given application.

To the best of our knowledge, only one survey focused on TP is available [61]: "Reviewing trip purpose imputation in GPS-based travel surveys". This work evaluates the TP from three perspectives, the first one is from the perspective of trip-end detection, to identify when a trip-chain is over. The second aspect evaluated is the feature selected to compose the model, which is evaluated in 4 different categories: geographical data, trip, and activity end, participant-related, and other data. Lastly, it discusses the algorithms and the validation. This work is focused on GPS travel surveys, discussing only techniques that could be applied for this type of dataset, while our work discusses the differences between all the datasets used in the literature.

We structure the remainder of this work as follows. In Section 2, we review the datasets available in the literature, discussing their potential and flaws. Besides that, we also discuss how this data can be enriched and potential areas that have been understudied and deserve a deeper analysis.

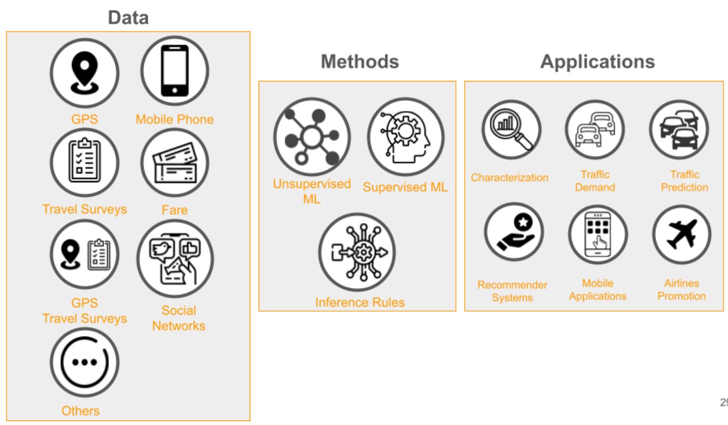

Fig. 1. Overview of the literature on TP

In Section 3, we review the methods used in the literature of TP, presenting their advantages and disadvantages, which tasks (single-trip, trip-chain), this technique tackles, and how it relates to each dataset available.

In Section 4 we discuss several applications that could benefit from the use of TP, pointing out works that used trip-purpose for specific intents. Lastly, in Section 5, we point out our conclusions and potential research questions for the TP classification task.

2 DATASETS

As we stated in section 1, proper data selection and modeling are essential for the task of predicting the purpose. Besides that, in most scenarios, a single dataset does not have all the features needed for this task, hence, a data enrichment process is a crucial step [58]. Thus, in this section, we make a deep review of the datasets used in the literature, pointing out their strengths and weaknesses, and sorting out different uses in the task of trip-purpose classification and trip-chain purpose classification. We also point to complementary datasets needed, and the process made in data enrichment.

The data in the Trip Purpose task can be summarized in two types, active and passive data. Active data refers to data directly collected from the user, for example, travel surveys and GPS travel surveys. In general, active data have explicit feedback from the user but suffer from the high cost of collecting such data. The passive data, which are collected without direct contact with the user, are collected in an automatized way. For example, we have Social Networks, GPS, and CDR data. Each of these datasets presents characteristics that make the data more or less suitable for different applications, as we discuss further in this section.

2.1 Travel Surveys

Travel surveys are conventional data collection methods used to gather information about individuals' travel behaviors and trip purposes [63]. These surveys involve direct interactions with respondents, often through questionnaires, interviews, or diaries. Before the use of computers, this process was a direct interview with the traveler using pen and paper [61]. These travel surveys are usually composed of 1-day or multiple-day travel diaries of sampled households. These households are selected according to the sociodemographic proportions of a population, being the aim to select a subsample that reflects similar travel behavior to all the population [15]. With time, pen and paper were replaced by computers and smartphones, making the data more reliable, and making it easier to analyze such data. While travel surveys have been widely employed for transportation planning and research, they possess both positive and negative attributes when utilized for trip purpose prediction.

Travel surveys provide a rich behavioral context for understanding the reasons behind trips [63]. Respondents can explicitly describe the purpose of their journeys, whether for work, leisure, or personal errands.

These surveys collect a wide range of information, including trip duration, mode of transportation, and sociodemographic details [53]. Since respondents provide information about their trips, the data can be highly accurate regarding trip purposes. This accuracy is particularly beneficial for validation and benchmarking purposes when evaluating predictive models [58].

Travel surveys are conducted with the explicit consent of participants, ensuring that privacy and ethical considerations are addressed. This is in contrast to certain forms of mobile and GPS data, which may raise privacy concerns.

Travel survey data is readily available through government agencies, research institutions, and transportation authorities. The accessibility, and the fact that this data presents an explicit target, make it a practical data source for TP classification.

Nevertheless, it is important to notice some disadvantages with travel survey datasets. Conducting travel surveys can be resource-intensive and time-consuming [15]. It requires significant effort in terms of data collection, recruitment, and participant engagement, resulting in high costs [15, 63]. As previously stated, the aim is to select a subsample that better represents the population, nevertheless, these studies suffer also from low response rates and long gaps between each study [53, 61]. For instance, in [53], authors aim to predict the trip purpose of long distance travels, nevertheless, the governmental long-distance travel dataset available is from 1995. Thus, for long-distance trips, there are not many datasets available, given the prohibitive cost of doing such a survey at a national level [53].

Due to the fact that these surveys are typically conducted at specific time intervals and may not capture detailed temporal or spatial aspects of travel behaviors, potentially limiting the precision of trip purpose prediction. Moreover, travel surveys often rely on a sample of the population, and this sample usually does not represent the entire population's travel behavior [15, 61]. Besides that, respondents may not recall their trips accurately, leading to memory biases in the data.

In general, self-administered surveys suffer from either under-reporting or over-reporting when describing a trip [7, 61, 70], especially due to the high effort to register all activities [58]. Hence, respondents might not provide truthful or complete information, introducing reporting biases that can affect the quality of the data. Thus, sample bias can lead to inaccuracies in predictive models. Furthermore, travel surveys are usually centered in small regions, being a small-scale dataset.

In general, travel surveys present a coarse granularity, especially when considering the aspects of mobility (displacement plus time) which are usually not available, and when available depend on the users' post-travel information, which as discussed, is not reliable [7]. If we consider the single-trip classification task, it is difficult to extract from the surveys each step of the travel. For the trip-chain classification, the lack of mobility features makes the task harder. Hence, for the task of TP, relying only on travel surveys is not enough.

Besides that, the data collected through travel surveys is typically gathered after the trips have occurred, making it challenging to use for real-time or dynamic trip purpose prediction (online prediction), which is critical for certain applications.

In general, travel surveys are much used in the enrichment process, given the fact that these are one of the few datasets available in the literature that presents an explicit TP indicated by the users [53]. Given this fact, authors use travel survey data with different sources to include a target feature in those datasets. For example, travel surveys are added to Governmental [53], and Mobility Platforms data [21].

In general, travel surveys can be seen as governmental data, given that most of the surveys used in the literature were conducted by governmental departments like in the United States [53, 55, 56, 90], and Japan [55, 56]. In this case, we refer to governmental data, data such as land use data, sociodemographic characteristics, economic development indicators, and travel and tourism statistics. In this scenario, the work of [53] uses individual and household sociodemographic characteristics, economic development indicators, and travel and tourism statistics to forecast the trip intention of single trips in long-distance scenarios.

Mobility platforms data as Google Maps, Google Places, Open Street Maps, have made the process of finding places to visit effortless since they present the places' characteristics and ratings. In this scenario, the work of [21] enriches travel surveys with data from Google Places characteristics as the categories of places nearby those visited by the users. Their results show that the use of data from Google Places can hence the quality of the proposed machine learning models.

In conclusion, travel surveys offer a wealth of behavioral context and comprehensive data for trip purpose prediction. However, they also come with limitations, including high costs, sample bias, and potential biases in self-reported data. Careful consideration of these positive and negative aspects is necessary when using travel survey data for transportation planning, research, and urban development. Balancing the strengths and limitations of travel surveys is crucial for accurate trip purpose prediction and the effective application of such data in various domains.

2.2 GPS

GPS (Global Positioning System) data has emerged as a valuable resource for a variety of data-driven applications. The widespread availability of GPS technology in various devices, including smartphones, vehicles, and wearables, means that GPS data can be collected from a diverse range of sources [58, 61]. Even though getting participants willing to share their location still is a hard task, GPS data is a readily available data source for research and practical applications. Besides that, different from travel surveys, GPS mitigates the burden on participants [61], providing more accurate spatial and temporal details, and avoiding problems such as under and over-reporting. As stated in [56, 61], GPS trajectories answers as "when" and "where", which motivated to infer the "why", that can be mapped in the trip purpose classification.

GPS data provides a high degree of precision and accuracy in terms of location information. It offers latitude and longitude coordinates, enabling a granular analysis of movement patterns and the identification of specific locations visited during a journey. This level of accuracy is pivotal for trip-purpose prediction.

Furthermore, GPS data often includes precise timestamps, allowing for the analysis of travel activities at different times of the day. This temporal granularity is valuable for distinguishing between various trip purposes based on the duration and timing of trips. GPS data can also capture additional geospatial information, such as altitude and speed. This supplementary data enriches the understanding of a person's travel behavior, facilitating the identification of the travel model, which is one factor that can help to understand a trip's purpose.

Despite its high data quality, GPS data can still be subject to inaccuracies and noise, particularly in areas with limited satellite visibility or urban environments with tall buildings [61], which are also known as urban canyons. Besides, managing and storing substantial volumes of GPS data can be challenging. Efficient data processing and storage solutions are necessary, particularly when dealing with large-scale GPS datasets [61].

The collection and use of GPS data may be subject to legal and regulatory restrictions [12]. Adherence to privacy and data protection laws is essential to avoid legal complications when using GPS data for predictive purposes. Hence, GPS data is hard to get and use, in special, it is hard to get users who are willing to share their location, and when using such data in more robust experiments, researchers use GPS-travel surveys, which mitigates some of the problems related to GPS. We discuss GPS travel surveys further in the next section.

It is noteworthy, that GPS data, on its own, may not provide sufficient context to understand the underlying reasons for trips, such as work-related, recreational, or personal errands [25, 58]. Hence, different from travel surveys, GPS datasets do not provide the Trip Purpose or any semantic information on the activities [12]. As stated in [12], in GPS data "...the trajectory data is rich, but activity information is poor".

Additional information, such as destination or activity type, is often needed to categorize trip purposes accurately. Hence, as mobile phone data, GPS data must be carefully scrutinized and manually classified to define the TP. Still, as pointed out in [58], the application of GPS data to infer TP is limited, especially in the recommendation area, given that peoples' geographical distribution does not reveal how much a user liked a venue.

Without the target, works define different proposals to have a labeled dataset. In [25], the authors propose a mobile application available in Switzerland that enables than to capture, with authorization, the users' mobility history. The author's main goal is to provide a TP prediction with as little information as possible. In this application, the users also label their travel, which makes easier the authors' work. In this work, authors state that the use of additional information (e.g. Google Places, Land use) did not have a benefit over their proposed methodology, thus, authors only use the information generated by their app. In this scenario, authors advocate against the use of external and personal information, given the limited data availability on a large scale.

Similar to the last work discussed, in [68], the authors propose a mobile application that collects data from participants of Rio de Janeiro. In this app, the users also label their travels. Their main goal is to provide in real-time the TP prediction. However, different from the last work, the authors have a small-scale dataset that is focused on a city, with few participants (19 users).

In [11, 12], authors use GPS data, provided by taxi cabs from New York, OpenStreet Maps, and also Foursquare data. The authors' main goal is to predict the trip purpose without the use of any devices. Their main hypothesis is that a region POI configuration indicates common human activities in a region. Besides that, the POIs' popularity plays a significant role in the region, and people do not move from one place to another to engage in the same type of activity. To accomplish so, first, authors define regions for each POI based on the city configuration retrieved from OpenStreets Maps. Then, Foursquare data is used to define for each region the POIs around the passenger dropoff. Besides that, the authors also use contextual information, as the origin point, and the time to define the TP to better estimate users' activity. Different from the other works, these two proposals do not have a ground truth. So, to measure the quality of their proposal, in [11], the authors use governmental data from a statistical regional sense to verify the accuracy of their prediction for a region.

In [42], authors also use taxi GPS and POI data to infer the TP. In this scenario, authors aggregate both of these data, creating areas of interest for the user. To validate their results, the authors extract the data from a Chinese taxi application, where the users manually label their travels. As we have seen in [11], the authors from [42], state that it is common for authors who use data from taxi data, to validate their results with travel survey data. Nevertheless, the authors discuss that in the scenario of taxi data, this is not reasonable due to the fact that travel surveys are usually for all different types of travel modes, and given the fact that the taxi data is in a finer granularity, while travel surveys are usually aggregate data.

In conclusion, even though GPS is a rich trajectory data, which can be extracted on a large scale, it lacks contextual information. In this scenario, if a manual label by the users or the researchers is not made, there is no ground truth to comparison. Hence, data enrichment is extremely needed to provide such information.

2.3 GPS Based Travel Surveys

GPS-based Travel Surveys, which are so known as GPS-assisted surveys [61], came to fill some gaps left by both GPS and Travel Surveys datasets. This data leverages the GPS technology to collect the users' positions and extracts explicit feedback from the users.

GPS-based travel surveys involve equipping participants with GPS-enabled devices, such as smartphones or GPS loggers. These devices continuously track the participants' location and record data points, including latitude, longitude, timestamp, and sometimes additional information like speed and altitude. This unobtrusive and passive data collection method eliminates the need for participants to manually record their travel activities, making it convenient and reducing the risk of data entry errors.

The process of extracting explicit feedback can be made in two different ways. First, users can answer questions related to the most recent trip, and make this at the end of each trip, or at the end of each period, like a diary [58]. In this process, the trips are segmented from mobile records, and then questions are sent to the users, asking for information about the trips [55]. Second, the GPS collects all the traces made by the user during a longer period (e.g. a week), and at the end, the user answers a questionnaire with queries related to their travel.

The collected GPS data is processed and analyzed to understand various aspects of travel behavior. This includes the frequency and duration of trips, the purpose of each trip (e.g., commuting, leisure, shopping), and the specific modes of transportation used (e.g., walking, cycling, driving, public transit). By recognizing patterns in the movement data, GPS surveys can automatically detect the mode of transportation during each trip based on movement patterns and speeds. This information is invaluable for spatial and temporal analyses, helping urban planners and transportation authorities gain insights into traffic congestion, transportation infrastructure usage, and travel patterns in specific areas. Also, as in GPS, this type of data provides a more realistic and continuous record when compared to travel surveys [55].

While GPS-based travel surveys offer a wealth of accurate and detailed data, these surveys have disadvantages similar to GPS and travel surveys. When considering aspects such as the scale these are usually focused on a small region, and usually in a small period of time [55], and also the fact that it is difficult to get people able to provide such information. Additionally, there may be costs associated with providing GPS devices to participants, and data volume needs to be addressed for effective survey implementation. Finally, the success of GPS surveys relies on participant compliance—ensuring participants consistently carry and use GPS devices for accurate data collection. However, GPS Travel surveys are exhaustive for the participants, given the fact that most of the records on the GPS devices are not logged in the diary, hence leading to under-reporting, given the amount of effort to register all activities [58].

As in travel survey datasets, we believe that GPS-travel surveys could not be used in an online prediction process. Given the fact that even though the GPS data can be available in real-time, the explicit feedback that is an integral and fundamental part of GPS-travel surveys cannot be extracted in real-time due to user constraints.

As in the travel surveys, GPS-travel surveys present the main advantage, the explicit travel purpose [90]. Besides that, it also presents the users' complete mobility trace which can be used to derive new features, and also to easily enrich the data. In this scenario, different from works that do not use any additional data with GPS travel-surveys [6, 55], some works enrich these GPS logs with Social Networks [15, 57, 58, 90], and Mobility Platforms [15, 56, 58]. In general, authors use data from other sources to get a view of regions, estimating the amount of place per category, and also the category visited by the users. Hence, an external view.

The work of [90] uses data from Foursquare, a Location Based Social-Network (LBSN), to augment the data from GPS travel surveys. In this scenario, the authors retrieved places near the points visited by the users in the survey. From these places were extracted their categories, and is measured the number of places in that region from each category. This process was called by the authors as semantic augmentation. A very similar process is repeated on other works, mostly changing the data source [15, 21, 56]. In [56], instead of Foursquare, authors used the Open Street Maps dataset, and in [15, 57, 58] authors used Twitter and Google Places for such estimation. Besides the methodology, in terms of data, in [57, 58], authors measure the nearby popularity of the POI to classify the TP.

One direction that we could observe in the literature is the preference for using data from Mobility Platforms to enhance GPS-based travel surveys because LBSN in general does not offer APIs to extract the data and the fact that, and LBSN platforms as Foursquare and Gowalla are not available anymore.

In conclusion, GPS-travel surveys bring benefits and disadvantages from both GPS and travel surveys. Even though it shows good results when evaluating trip purposes, especially because it has explicit feedback and user mobility traces, the authors show that the use of external data is essential to achieve better results. However, the data augmentation process can be very costly, especially when using paid APIs such as Google Places [25]. In our evaluation, it was possible to notice that most of the works in the TP literature use GPS-travel surveys, since the data already has the travelers' context, can be easily enriched, is easy to validate (given that it has the labels), and has the users' mobility. Nevertheless, we high-value that these are small-scale datasets, with evaluations made in small regions during a small period, thus, dependent on the application proposal, as in RS, this type of data cannot be applied given the different mobility patterns by regions and countries [67].

2.4 Mobile Phone

Mobile phone data has emerged as a valuable resource for a range of data-driven applications and applied to different mobility studies [1, 27], including trip purpose prediction. Mobile phone data offers a wealth of advantages for this task, but it also comes with its own set of limitations.

In the literature, three different types of mobile phone data have been used for trip purpose prediction, Call Record Detail (CDR), Mobile Signaling Data (MSD), and Cellular Network Data (CND). CDR is characterized by storing actions (e.g. receive/make calls, receive/send texts) made by a cellphone company user. Between the data store, there is the action' caller and recipient ID, type of action, but most importantly for this context, timestamp, and the users' location [8]. It is noteworthy, however, that in CDR data, we only have the geo-coordinates for the tower that the user was connected to while making the action, and not the exact user's location, hence, it is a coarse granularity. As in CDR, MSD only has the geo-coordinates of the tower, the difference is that it has all the time that a user was connected to a cell, hence is a finer granularity than CDR. Different from CDR and MSD, CND works like GPS data, saving the users' position every few seconds [45, 88], thus, this type of data presents the same advantages and disadvantages as GPS.

Despite its potential, mobile phone data also comes with several limitations for trip-purpose prediction. First, data accessibility is a limitation, as comprehensive mobile phone datasets may not always be readily available for researchers and practitioners. These are usually limited to the mobile phone companies. Besides that, data quality and noise are significant concerns. Mobile phone data may suffer from inaccuracies caused by issues like signal dropouts, GPS inaccuracies, and incorrect cell tower associations [8]. Such errors can lead to misclassification if the dataset is not on a large scale. Moreover, these types of datasets usually do not contain any contextual information, in special, a labeled trip purpose. In these scenarios, authors usually consider unsupervised machine learning approaches to define the purposes, but we discuss this further in the Techniques section ??.

The works presented in [45, 88] use data from CND collected from Beijing to predict Trip Purpose. To enrich the data, having contextual information, the authors collected information from Baidu Map, extracting the categories of POI near the users' destination. In this scenario, to validate their results, the work [45], compares their results with the Beijing household census. It is noteworthy, however, that the user only compares their results in terms of the amount of travel classified in each category, and does not consider the region. As stated in [12], the region plays an important role in the definition of purpose, hence, this type of comparison should be made by region.

In [71], the main goal of the authors is to identify the trip purpose of tourists in the city of Xiamen, China. To accomplish so, the authors use mobile signaling data, and data from a "city electronic map" (not specified) to extract POI. Lastly, they survey the people they extract the coordinates to label their data.

As stated before, CND has the same advantages and disadvantages as GPS data for the task of TP classification, nevertheless, the use of CDR data is even more challenging given the coarse granularity. Given this additional difficulty, very few works use CDR data in their evaluation.

In [1], authors use CDR and land-use data to characterize the Trip Purpose. The data is collected from the Boston metropolitan area. In this scenario, authors verify the users' mobility patterns, considering the time to define areas such as home and work. Other places that are not defined as home or work as put in this third, generic, category. To validate their characterization, the authors compared the results with different travel surveys from the Boston region. It is noteworthy, however, that predicting and defining their users' homes and work, is considered in general an easier task [45], than predicting the other categories, given that fact that you just have to consider the users' mobility patterns when these are available in a fine granularity, or more regular patterns [27].

In conclusion, mobile phone data holds significant promise for trip-purpose prediction due to its richness and temporal granularity. However, addressing challenges related to data quality, privacy, behavioral context, and data integration is essential to fully leverage its potential. Considering the applications for mobile phone data, depending on its granularity can be even harder to properly classify a trip. In this scenario, CDR data, as used in [1] is better suited for characterizations and urban planning evaluations, given its coarse granularity it is not possible to define the place that the user is. Besides that, this type of data can be used in applications that sum up collective behavior, such as traffic prediction.

2.5 Social Networks

Social Networks (SN) enable users to share their experiences with other people through social networks and traveling specific platforms [80]. The popularization of these platforms made studying tourism a simpler task given the amount of data available [2, 40]. There are currently available in the literature datasets extracted from Location-Based Social Networks (LBSNs) such as Gowalla, Foursquare, Twitter, and Yelp, and Travel Social Networks (TSNs) such as TripAdvisor, Real Travel, and Travello.

SN made it possible to seek Points of Interest (POI), provide and receive information about places, and share experiences from a unique point of view. Studies show that 80% of American travelers use SNs while traveling, and more than half of this percentage share information of their journey with their contacts [80]. These studies also show how SNs are the tool most adopted by travelers [80].

Between these platforms, three different types of SNs are generally used: General-Purpose SN, Location-Based SN (LBSN), and Travel SN (TSN). General-purpose SNs as Facebook and Instagram, usually contain information on a variety of topics. This information can be geotagged, which can enable to verification user's mobility and activity purpose. However, these are usually restricted to the user's social circle and are not structured to easily identify the users' opinions and purposes.

LBSNs are social platforms where users share in real-time their geospatial location and visit timestamp, composing a visit check-in [4]. The users of these platforms show a pattern of use on a regular daily basis, and about 90% of their transitions between places are within the 50 km range [50]. Corroborating with that, the work of [28] shows that most of the locations are followed by fewer than 5 locations consecutively, showing that users have a similar visitation pattern, and follow the human mobility patterns found in CDR datasets [27].

Considering LBSN datasets available in the literature, two different sets of features can be found. Datasets from Foursquare and Gowalla present a small set of features, composed of the User identifier, POI identifier, POI category, and Visit Timestamp. In the case of Foursquare, it was possible to retrieve the users' opinions of a place through a review on Twitter. Nevertheless, as in general-purpose SNs this data is not structured, and to analyze how much a user liked a place, it would need a sentiment analysis step. In this case, works that employ Foursquare and Gowalla data assume that if a user visited a place, then, the user has liked, since no explicit feedback is available [37, 82, 83]. On the other hand, Yelp data is more suited for feature extraction, given that besides having the check-in data, it enables the users to share an explicit evaluation for each POI.

As we have seen in the previous discussion, the usage of social networks' datasets for the task of TP, in special, LBSN' Twitter has been used to enrich data from other sources [15, 57, 58], but never as the main data. This usage occurs especially in scenarios where the data is not labeled, and the SN' data is used to provide contextual/semantic information. In these scenarios, the usage of Twitter (currently X) was mostly motivated by easy access to the data through APIs [61].

When considering the TP task, there are a lot of advantages to using LBSN data in the model. First, different from GPS Travel Surveys available in the literature, LBSN is on a large scale, enabling mobility check-ins in any part of the world. Besides that, it also presents explicit feedback, even though it is not structured, that enables one to understand how much a user likes a place, and who is with the user, and in some scenarios, the textual features can even explain the visitation purpose. Nevertheless, to the best of our knowledge, few works make use of textual features [36, 48, 64, 74], and when used, the data used is the characteristics from a place (or a set of places) and not a review made by the user. Considering these facts, we believe that this is a hot topic that can be explored, and used to enhance the quality of models.

Nevertheless, when dealing with LBSN, lots of challenges can appear. First, different from GPS and CDR data, the check-ins on LBSN are dependent on the users' will to make the check-in, hence, been in most cases more sparse than CDR data. Hence, making a trip chain to make the evaluation can be a challenge, since the time between check-ins of the same user can be very sparse. Thus, defining when a trip ended is also a difficult task that has been studied [19, 61]. Besides that, LBSN data does not present the ground truth, hence, to have labeled instances, authors use automatic techniques to label their datasets [48, 58], or manually label a small set of instances, to train supervised models.

Besides LBSNs, another type, TSN data can also be used for TP evaluation. TSNs are social platforms focused on tourism, where users can share places visited and reviews with other travelers. Different from LBSNs, TSNs focus on a "couch review", i.e. the travelers make their review after visiting a place and in many cases months after that visit. In this scenario, the check-in time is not available, only the month and year of the visit. However, the information presented in a TSN review is more detailed, given that the user supplies explicit ratings to aspects such as place service, cleanness, location, and trip purpose, also providing a written review. The main advantage of TSN datasets over Yelp data is that place owners also share their profile, defining features such as price, and opening hours, and having a closer relationship with the travelers, hence, more contextual data.

This also shows the difference in the behavior of users from LBSN and TSN. While LBSN makes reviews on daily activities (e.g. car shop, grocery stores, restaurants), hence, closer to their homes (50 KM range), TSN users' reviews are more focused on travel experiences, and are more directed to interesting places.

Besides the use of TSNs/LBSNs data, it is also common in the literature the use of different datasets in tourism-related works. In some cases, external datasets are used to increase the amount of data and enhance the quality of data, making it more suited for real scenarios [35].

Using social media information enables an improvement in detection but with a lack of robustness. When driving, attending classes, or working, people are less likely to tweet compared to when going shopping or eating out [61]. Social media data would be biased toward young people and against the elderly. Data from LBSN is noisy, and very hard to extract relevant information [58].

To the best of our knowledge, only one work uses SN as the main dataset for TP task. In [62], authors use textual data scrapped from TripAdvisor to define the Trip Purpose. To label their data, the authors extract keywords from the reviews made by users and assign to each review one of the eight labels defined by the United Nations World Tourism Organization. Nevertheless, the main purpose of this paper is not to assign to each review a trip purpose, but to use this information to enhance the quality of recommendations for travel destinations. This way, as the authors' focus is not on the TP, they do not discuss the quality of these results, but it show one potential usage of TP information.

In conclusion, the usage of SN has shown great benefits when enriching other data sources [55], but has been understudied as the main data source for the task of TP. For real-time applications, such as RS and traffic prediction, SN can be used given the volume of data generated. Nevertheless, deep characterizations for urban planning the use of SN can be difficult given that the users are not willing to share daily activities (like driving). One great opportunity that appears is the use of textual features to extract the trip purpose. To the best of our knowledge, Natural Language Processing (NLP) techniques like text classification have not been used to classify trip purposes.

Lastly, in Table 1 we present a summary of this section, presenting all the datasets discussed, their advantages, and disadvantages. We also present how these datasets are collected, if they are annotated or not, which type of task this dataset can be used, the granularity, and to which these datasets are enriched with.

Dataset	Collected	Annotated	Single Trip Purpose	Trip Chain Purpose	Granularity	Online task	Enriched with	Advantage	Disadvantage
Travel Survey	Actively	Yes		x	Coarse-Grained	No	-	Explicit feedback Explicit target	Self-administred survey Small scale Cannot check mobility
GPS	Passively	Can be	x	x	Fine-Grained	Yes	LBSN Human Mobility Platforms Governmental data Travel Surveys	Mobility trace	Small scale Accessibility
GPS Travel Surveys	Both	Yes	x	x	Fine-Grained	No	LBSN Human Mobility Platforms Governmental data Travel Surveys	Mobility trace Explicit target	Small scale Self-administred survey
Mobile Phone	Passively	No	x	x	Vary	Yes	LBSN Human Mobility Platforms Governmental data	Mobility trace	Small scale Fixed to cell towers Accessibility
LBSN	Passively	No	x	x	Fine-Grained	Yes	Human Mobility Platforms	Large Scale	Depends on users making a voluntary check-in
TSN	Passively	Yes	x	x	Coarse-Grained	No	-	Large Scale Explicit target	Depends on users making a voluntary check-in Cannot check mobility

Table 1. Overview of the datasets used in the TP literature

3 METHODS

Given the understanding of the data used for TP, selecting the suitable technique is the next step. There are different techniques have been applied in the task of TP classification. These techniques can be classified into three different categories (*i*) Inference Rules, (*ii*) Unsupervised Machine Learning, (*iii*) Supervised Machine Learning. It is noteworthy that certain techniques can only be applied to certain data, so in this section we describe these scenarios, discussing their advantages and disadvantages, and the more suitable applications.

3.1 Inference Rules

Inference rules are logical guidelines used to classify the purpose of travel based on available information. This is a simple technique that observes patterns, and through them defines an outcome. These patterns can be defined in a probabilistic way, or through rules authors' previous definitions.

Considering the TP scenario, few works use inference rules to define the intention [1, 6, 26, 66]. Usually, these are statistical rules that must be revised with time, given that these do not consider the constant change in human mobility. These rules are based on patterns of place the user spends most of his/her time at home, and the place where the user is during business hours as work, being other activities categorized in a generic group [1]. Considering characterization scenarios, and aggregated studies in more simplistic scenarios, these rules can be applied. Illustrating that, these could be used in urban planning, because it gives a good idea of the users' activity mobility. Besides that, this technique has a small computational time, which allows the use it online applications, and can be easily applied to any dataset. Nevertheless, this technique fails to capture the users' purpose in a finer granularity (i.e. other trip purposes besides work and home), also having problems defining users that do not follow the usual circadian cycle, working at night, or home workers.

In conclusion, inference rules are a simple and interpretable technique that can be applied to any scenario, with a cost that will not provide a finer granularity of prediction. In addition, these rules must be constantly revised given the dynamics change. These techniques can work well for more static, and daily activities, nevertheless, will not work for tourists, and people who do not follow the established pattern.

3.2 Unsupervised Machine Learning

Using unsupervised machine learning to classify trip purposes involves the application of data-driven methods that do not rely on predefined labels or categories. Instead, these algorithms identify patterns and group data points into clusters based on inherent similarities. These techniques are applied in several scenarios, given their high applicability with raw data, and generally good results. One difficulty with this type of technique is the fact that after the application, an interpretation of the defined groups must be made, which can also change depending on the scenario.

Considering the context of TP, this technique has been applied in several works, given the fact there is no need for labeled data, which is hard to get. Thus, this makes this type of technique suitable for every scenario. In [44], the authors discuss that supervised ML techniques require additional data collection (e.g. contextual data) to accurately infer the trip purpose. Complementarity, in [12], authors argue that unsupervised ML is the most suitable technique for TP classification given the lack of large-scale labeled data. Besides that, when the label is available, the ground truth is measured based on the users' recalling of their travels, where the users cannot remember correctly what they have done, as we discuss for travel surveys. However, this type of technique usually presents a difficulty related to the validation of their proposal, given the lack of labeled data. To validate their methods, authors search travel surveys, and land-use data to identify regions that are related to the activities characterized. However, as discussed in [42], travel surveys are not suitable for every scenario, especially those where the data is directly extracted from one vehicle (e.g. taxis, bikes, scooters), since travel surveys represent aggregated data from different sources.

One difficulty when using unsupervised ML is how to properly model the data. In the literature, it is possible to notice three different approaches for such modeling, namely, Topic-Modeling [64], Markov-Chain [44], and Clustering [12] techniques, each one presenting its own sets of particularities. It is noteworthy, that each technique will provide its evaluation on its granularity, for example, analysis using topic-modeling can only be made at a group level, while clustering techniques allow grouping the trips at an individual level. We discuss this further in the following sections.

Topic-Modeling. Topic Modeling (TM) addresses the problem of discovering relations between Documents (D) and Topics (Z), and also takes into consideration the relation between the Terms (W). This type of technique allows classifying, retrieving, organization, and analyzing a document [32].

Currently, Topic Modeling has been widely applied in the automatic evaluation of vast text quantities, making it possible to describe texts through his capacity to summarize documents and represent them through topics. These models work by simultaneously performing dimension reduction and clustering [54]. So, unsupervised models, without any previous knowledge, find cohesive linked topics that summarize the document. Thus, given an amount of K topics in a set of documents D, the algorithm will perform by finding topics that better describe a document $d \in D$ in a time slice $t \in T$, allowing also to analyze the evolution of a topic, given that most of the proposal performs by using chained distributions, hence, a growth and decrease of a topic is linked with it evaluation in a previously time slice.

As we can see, this type of technique was initially designed to reduce and summarize textual data and has been adapted to work with TP [46]. In this scenario, each document is a trip, each term is a trip feature (e.g. origin, destination, and time), and each topic is a cluster that describes the trip' purpose.

Even though there are different techniques for topic modeling, such as Non-negative Matrix Factorization (NMF), Singular-Value Decomposition (SVD), and Latent Dirichlet Allocation (LDA), most of the works only use LDA as technique [30, 34, 45, 64, 74, 88]. According to [45], LDA can automatically uncover patterns from raw data, which can be instrumental for TP inference. It is noteworthy, however, that the LDA is not a flexible technique [45]. So, to properly use it in the TP context, improvements must be made. Potential future research could address different techniques in topic modeling for the task TP by directly comparing them. Besides that, there are techniques for topic modeling that take into consideration the temporal aspects of the topics [33], which can be a factor that allows a more precise evaluation of the TP context given that this is a factor that directly impacts mobility.

As previously stated, one downside of this type of technique is that after defining each topic (trip purpose), authors must manually interpret each topic and try to put a label (category), and each one, which can be time-consuming. Another point about these techniques is that each trip can contribute to the construction of more than one topic. Hence, it is not possible to separate with confidence each topic a trip belongs to. Besides that, this is a costly technique that does not allow partial fit, so for adding new trips to one category, new training must be done. Given these facts, this type of technique can only be applied to scenarios where group evaluations are made, given that we could put each trip in just one topic.

Markov-Chain. Markov chains can be used to classify trip purposes by modeling the sequential nature of travel behaviors and decisions. Markov chains are mathematical models that describe a system's transitions between different states, where the probability of transitioning from one state to another depends only on the current state. In the TP context, each state represents a potential trip purpose. For example, states could include "leisure," "business, "family vacation," and so on.

This type of technique works better with datasets with finer granularity, given that the state will be better represented like GPS, and GPS-travel surveys [44**?**]. Nevertheless, even when using GPS-travel surveys, the additional (contextual), may not be used. But in this scenario, it is easier to define the states, given that these must be previously defined. The main advantage of Markov-chain is that it can be used to model both single-trip and trip-chain and in both cases, it does not need many adaptations. Nevertheless, to properly work, Markov chains need a huge amount of data.

Clustering. Unsupervised machine learning clustering techniques are used to group similar data points without any predefined labels or categories. In this scenario, the similarity between the points inter-cluster is bigger than the similarity between the points extra-cluster.

Considering the trip-purpose task, this type of technique works well with raw data and works for both single-trip and trip-chain. When dealing only single-trip or trip-chain it is possible to use directly raw data in the algorithms, made in [3], in which authors use DBScan, and [22] of uses Agglomerative Hierarchical Clustering. However, to properly work for both scenarios, a pre-processing step must be made to put the data in the same n-dimensional space. To accomplish this, authors use embedding techniques [12] which are much used in NLP to put data textual data in the same space, while maintaining the semantic data. After doing such pre-processing, authors apply clustering techniques as k-means [12]

In sum, this type of technique works well with raw data, and with the specific pre-processing can be used to group both single-trip and trip-chain. Clustering techniques can also be used online, and are recommended in scenarios where there is no label available, which is most of the scenarios, performing well with proper modeling in all the scenarios. The biggest downside of this type of data is for the evaluation of the models, given that most work compares their achievements with travel surveys and governmental data, however, this cannot be enough in some scenarios.

3.3 Supervised Machine Learning

Using supervised machine learning to classify trip purposes involves training a model with labeled data, where each trip is associated with a known purpose (e.g., business, leisure). The model learns from this labeled data to make predictions on new, unlabeled trips.

Considering the TP scenario, this type of technique is limited to datasets that have the label available, like travel surveys, GPS travel surveys, or specific types of GPS and social networks. Even though, this type of technique can be dynamically retrained and used online, and it is easily interpretable, it needs lots of contextual data to properly identify each class. Hence, this needs a deeper pre-processing and enrichment, which can be costly steps [25].

Another problem that is not usually addressed by researchers is the unbalanced purposes of the training. This is a common scenario, given that people tend to make more trips with one purpose, like going to eat, than other purposes, like going hiking. In scenarios like this, the typical accuracy evaluations are not suited given that the model must have a good assertiveness for all classes, and not only the majority class [31]. However, only a few works tackle this problem. In this scenario, the proposal of [25] tackles this problem by using an oversampling technique for the majority class. Another technique that can be used, is to make an instance selection of the majority class, reducing the amount of training instances.

In the literature, different classification models have been applied for TP classification tasks, like Decision Trees [16, 39, 53, 86], Random Forest [21, 25, 59], SVM [90], Bayes-Theorem [11, 17], Neural-Networks [15, 48, 55, 56, 76], each of these techniques has its own particularities, that are better-discussed in [38].

Models based on trees (e.g. Decision Trees, and Random Forests), present as main advantage of their interpretability, and the fact that they can be easily associated with inference rules, being a natural evaluation step. The same goes for the Bayesian models. SVM and Neural Networks are more robust techniques, that can achieve better results, but at the cost of computational expense and lack of interpretability. But both these models handle better unbalanced data, which assists in a scenario so unbalanced as trip-purpose.

In conclusion, even though, supervised ML is vastly applied in TP classification, it comes with the disadvantages of needing labeled data, which is hard to get, and making a good pre-processing for the model to work properly. Besides that, properly selecting the right model for the task can be a difficult task. What researchers usually do is to compare several techniques and select the one that maximizes the model evaluation metric. However, if the data is available this type of technique is more suitable for the task, given that it can be easily put online, making it suitable for real-time applications like recommender systems, and traffic prediction. Besides that, this technique can be more easily validated, and interpreted, with no need for a posterior evaluation of the results.

Finally, each of these techniques has its own sets of advantages and limitations, as it was discussed inference rules are a simple static technique that can be used in simpler scenarios like daily activities' characterization. Unsupervised ML models present different modeling that can vary according to the task that is being tackled, besides that, depending on the granularity that the answer will be served (i.e. individual or grouped) the technique can also vary. The great advantage of such modeling is that it does not need a label and works well with raw data. Lastly, the supervised ML model needs labeled data to be trained. This can be an issue, given that most large-scale datasets are collected passively, without direct contact with the user (GPS and CDR). However, when the label is available this is a simple technique to evaluate its quality, and also can be easily applied to real-time applications. The Table 2 gives and overview of everything discussed in this section.

Technique	Needs Label	Online	Works well with	Task	Classification level	Applications	Advantages	Disadvantages
Inference Rules	No	Yes	GPS, CDR	Single-Trip	Individual	Simple characterization Urban planning	Simple technique Interpretable No need for label data	Low confidence in dynamic scenarios
Topic Modeling	No	No	Any type of data	Single-Trip/Trip-Chain	Group	Urban planning Characterization	No need for label data	Can need post procesing Posterior interpretation Computationally costly Hard to validate
Markov Chains	No	Yes	GPS	Single-Trip/Trip-Chain	Individual	RS Characterization Urban planning Traffic prediction	No need for label data	Computationally costly Hard to validate
Clustering	No	Yes	Any type of data	Single-Trip/Trip-Chain	Individual	RS Characterization Urban planning Traffic prediction	No need for label data Can be interpretable	Can need post procesing Posterior interpretation Computationally costly Hard to validate
Supervised ML	Yes	Yes	Any type of data	Single-Trip/Trip-Chain	Individual	RS Characterization Urban planning Traffic prediction	Easy to validate Can be interpretable	Traning is costly Needs labeled data

Table 2. Overview of the techniques used in the TP literature

4 APPLICATIONS

4.1 Characterization

Data characterization refers to the process of understanding and describing the essential qualities and attributes of a dataset. It involves examining the structure, patterns, relationships, and statistical properties within the data to gain insights into its nature. This process is crucial for comprehending the content, quality, and potential uses of the data. With a good data characterization process, it is possible to make a data-driven decision-making process, giving more confidence in the process [29].

Considering the TP context, a deep data characterization can help in several scenarios, from human mobility, epidemiology, and urban planning (which is deeply discussed further) [1]. The characterization opens the door for opportunities that can be found in the data. Illustrating that, the work [1] makes a characterization of a mobility dataset, identifying users' patterns in activities like home and work. With a more robust methodology, in [3] authors make a characterization of human activity intention from GPS data sources.

In conclusion, characterization techniques can be applied to any data source and can be used both to discover intention patterns and describe these patterns.

4.2 Urban Planning

Urban planning is the discipline and practice of designing, managing, and shaping the physical environment of cities and communities [41]. It involves making decisions about how to use land, resources, infrastructure, and public spaces to create functional, sustainable, and attractive urban areas.

Urban planners consider various aspects such as transportation, housing, public amenities, environmental sustainability, economic development, and social equity when developing plans for a city or region [78]. They aim to improve the quality of life for residents by creating environments that are efficient, safe, accessible, and aesthetically pleasing.

These professionals work with government agencies, developers, communities, and other stakeholders to develop long-term plans, policies, and strategies that guide the growth and development of cities while taking into account factors like population growth, cultural diversity, economic changes, and environmental concerns. Urban planning plays a crucial role in shaping the way cities evolve and function, striving to create spaces that are both livable and sustainable for generations to come.

More recently, urban planners have benefited from the digitalization of travel surveys, from pen and paper to electronic surveys, which allowed a faster and more accurate data evaluation [61]. Besides that, the popularization of Geographic Information Systems (GIS), and mobility platforms, allied with travel surveys, allowed these professionals to enhance the scale of their evaluations, providing a more detailed analysis of city dynamics [1].

Considering the context of the trip purpose, it is well known that urban planners benefit from such information to estimate the size of the streets, public transportation lines that will attend a neighborhood, land use allocation, and service provision [72].

In the end, for such application, it is common for urban planners to use travel surveys and census data, given that these are less biased toward a population group. However, as discussed, these professionals can benefit from other sources, however, when making such analysis it is important to understand the data limitations and biases.

4.3 Traffic Prediction

Traffic prediction involves using historical and real-time traffic data to forecast future traffic conditions. This problem can be formulated as: let T represent time, and $V(T)$ represent the traffic volume at time T, so the prediction formula can describe as:

$$V(T) = f(V(T-1), V(T-2), \ldots, V(T-n)) + \epsilon. \tag{1}$$

As we are discussing throughout this paper, contextual data plays a huge role in mobility analytics. Thus, the use of different features as events, trajectories, types of roads, and temporal information [60, 81]. In [14], authors verify the impacts of weather and trip purpose on the traffic, showing how these two can deeply impact the population behavior. Nevertheless, to the best of our knowledge, no model of traffic prediction takes into consideration the trip purpose to enhance their qualities. Hence, this is a potential work that could be tackled in the future.

4.4 Recommender Systems

Recommender Systems (RS) are algorithms developed to provide to a user through the learning of historical data personalized suggestions [52]. These algorithms tackle the problem known as information overload [89] or choice paralysis [24], by suggesting the most suitable items to specific users, predicting the consumer interest [52]. In our context, also known as POI recommendation, the RS is responsible for predicting to each place p not visited by the user u a rating ($\hat{y}_p$) that summarizes how much u will like that place.

Thus, POI recommendation is the task of suggesting new places for users to visit based on their history. Since the use of LBSN has popularized, the interest of academics and industry has increased further [50, 87]. As stated in [4], "location data bridges the gap between the physical and digital world and enables a deeper understanding of users' preferences and behavior".

Different from conventional recommendation scenarios (e.g. movie recommendation), POI recommendation is considered to be a harder task, due to geographical (e.g. physical constraints between the POIs) [10], social (e.g. friends can influence on the user visit) [4, 87] and temporal (e.g. places are more searched on summer) influences [50]. Besides, the number of instances tends to be much smaller, sparser, and noisier than in book and movie scenarios [5, 79]. Also, the heterogeneity of information in social networks describes the user activity from a variety of perspectives and through different types of data (e.g. photos, text, check-ins), which leads to a vast literature that focuses on different approaches for accomplishing the POI recommendation task [13, 23, 85].

First works on POI recommendation system, focused on recommending a top k list of locations to be chosen by the user [13, 77, 85]. Due to the many aspects that make the POI recommendation scenario harder, works propose methodologies based on the use of additional information. As stated in [73], additional data allows for more accurate recommendations than traditional methods, thus, works exploit geo-coordenates [13, 47], social ties [77, 85], POI category [43, 84] and time [85] to enhance predictions quality.

It is noticeable that spatial and temporal are between the covered features within these models, however, few works use the trip intention to provide suggestions [51, 62]. However, it is noteworthy that considering the trip purpose while recommending a place is essential to provide a suitable suggestion given the users' context. For example, on a business trip, it is better to recommend places that are near the hotel where the user is staying, while on leisure trips, the user may want to visit places throughout the town.

Given the difficulty of extracting the trip purpose, there are only a few works that use it. In both papers [51, 62], authors use data from SN due to the heterogeneity of information available, making this source more suitable for recommendation scenarios. As discussed earlier, except for travel surveys and GPS travel surveys, in natural conditions all other sources do not present any type of explicit user feedback, thus making it difficult for RS to properly identify the users' opinions about one place to make recommendations. Hence, SN is the most suited data for this type of task given its scale, with lots of users' historical data, and explicit feedback allowing to proper verify the users' opinions about places and things.

4.5 Companies

Lastly, during our research, we could see opportunities for the use of TP outside the academy. Illustrating that, the work of [49], uses airline company data to estimate the amount of business and leisure travel, enabling the company to offer these travelers promotions and better service. Besides that, restaurants and attractions, could benefit from such data to also provide promotions for other companies. Illustrating that different services can be offered for people who want to have a business meeting in a restaurant. For attractions, they could provide discounts for groups of travelers. Besides, these companies could benefit of understanding their customers' behavior, to better estimate cost, maximizing their gain.

In conclusion, have several applications inside and outside the academy. As we saw, TP can be used within the government to design better cities, enhancing their people's lives' quality, and can be used to enhance the quality of travel prediction models, and recommender systems, providing a deeper personalization and increasing the quality of suggestions. Besides that, this type of information can be used outside the academy to provide promotions and attract more customers to business.

5 CONCLUSIONS

Trip purpose and contextual information are essential to applications in mobility but have been understudied in this scenario. Even though this can be applied in several scenarios, from recommender systems to traffic prediction, this is a hard task due to the ambiguous nature of human mobility. Besides that, properly identifying datasets and techniques to be applied in an application can be a difficult task, given that, the methods and datasets must be tailored for regions given cultural boundaries, and the ever-changing mobility scenario.

This work surveyed the research efforts for the trip purpose classification task. We introduced basic concepts for differentiating single-trip and trip-chain. For each of the tasks we introduce possible applications, datasets, and methods data can be used. We discuss this, providing researchers and practitioners with clear guidelines to understand what is the most suited dataset and technique given the application for the Trip Purpose. Finally, in each section, we present in discuss possible opportunities for problems that are still open in the literature and can be tackled in the future.

REFERENCES

[1] Lauren Alexander, Shan Jiang, Mikel Murga, and Marta C González. 2015. Origin–destination trips by purpose and time of day inferred from mobile phone data. *Transportation research part c: emerging technologies* 58 (2015), 240–250.

[2] Francisco Amaral, Teresa Tiago, Flávio Tiago, and Androniki Kavoura. 2017. Comentários no TripAdvisor: Do que falam os turistas? *Dos Algarves: A Multidisciplinary e-Journal* 2, 26 (2017), 47–67.

[3] Ali Arian, Alireza Ermagun, and Yi-Chang Chiu. 2017. Characterizing activity patterns using co-clustering and user-activity network. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 1–6.

[4] Jie Bao, Yu Zheng, David Wilkie, and Mohamed Mokbel. 2015. Recommendations in location-based social networks: a survey. *GeoInformatica* 19, 3 (2015), 525–565.

[5] Punam Bedi, Sumit Kumar Agarwal, Vinita Jindal, et al. 2014. Marst: Multi-agent recommender system for e-tourism using reputation based collaborative filtering. In *International Workshop on Databases in Networked Information Systems*. Springer, 189–201.

[6] Wendy Bohte and Kees Maat. 2009. Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands. *Transportation Research Part C: Emerging Technologies* 17, 3 (2009), 285–297.

[7] Stacey G Bricka, Sudeshna Sen, Rajesh Paleti, and Chandra R Bhat. 2012. An analysis of the factors influencing differences in survey-reported and GPS-recorded trips. *Transportation research part C: emerging technologies* 21, 1 (2012), 67–88.

[8] Francesco Calabrese, Laura Ferrari, and Vincent D Blondel. 2014. Urban sensing using mobile phone network data: a survey of research. *Acm computing surveys (csur)* 47, 2 (2014), 1–20.

[9] California. 2019. California Household Survey. https://dot.ca.gov/programs/transportation-planning/division-of-transportation-planning/data-analytics-services/transportation-economics/ca-household-travel-survey. Accessed: 2023-10-17.

[10] Buru Chang, Yookyung Koh, Donghyeon Park, and Jaewoo Kang. 2020. Content-Aware Successive Point-of-Interest Recommendation. In *Proceedings of the 2020 SIAM International Conference on Data Mining*. SIAM, 100–108.

[11] Chao Chen, Shuhai Jiao, Shu Zhang, Weichen Liu, Liang Feng, and Yasha Wang. 2018. TripImputor: Real-time imputing taxi trip purpose leveraging multi-sourced urban data. *IEEE Transactions on Intelligent Transportation Systems* 19, 10 (2018), 3292–3304.

[12] Chao Chen, Chengwu Liao, Xuefeng Xie, Yasha Wang, and Junfeng Zhao. 2019. Trip2Vec: a deep embedding approach for clustering and profiling taxi trip purposes. *Personal and Ubiquitous Computing* 23 (2019), 53–66.

[13] Chen Cheng, Haiqin Yang, Irwin King, and Michael Lyu. 2012. Fused matrix factorization with geographical and social influence in location-based social networks. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 26.

[14] Mario Cools, Elke Moons, Lieve Creemers, and Geert Wets. 2010. Changes in travel behavior in response to weather conditions: do type of weather and trip purpose matter? *Transportation Research Record* 2157, 1 (2010), 22–28.

[15] Yu Cui, Chuishi Meng, Qing He, and Jing Gao. 2018. Forecasting current and next trip purpose with social media data and Google places. *Transportation Research Part C: Emerging Technologies* 97 (2018), 159–174.

[16] Zhongwei Deng and Minhe Ji. 2010. Deriving rules for trip purpose identification from GPS travel survey data and land use data: A machine learning approach. In *Traffic and Transportation Studies 2010*. 768–777.

[17] D Dhananjaya and T Sivakumar. 2022. A trip purpose inference framework using spatial clustering and bayesian probability. (2022).

[18] Linus W Dietz, Rinita Roy, and Wolfgang Wörndl. 2019. Characterisation of traveller types using check-in data from location-based social networks. In *Information and Communication Technologies in Tourism 2019: Proceedings of the International Conference in Nicosia, Cyprus, January 30–February 1, 2019*. Springer, 15–26.

[19] Jianhe Du and Lisa Aultman-Hall. 2007. Increasing the accuracy of trip rate information from passive multi-day GPS travel datasets: Automatic trip end identification issues. *Transportation Research Part A: Policy and Practice* 41, 3 (2007), 220–232.

[20] Erik Elldér. 2014. Residential location and daily travel distances: The influence of trip purpose. *Journal of Transport Geography* 34 (2014), 121–130.

[21] Alireza Ermagun, Yingling Fan, Julian Wolfson, Gediminas Adomavicius, and Kirti Das. 2017. Real-time trip purpose prediction using online location-based search and discovery services. *Transportation Research Part C: Emerging Technologies* 77 (2017), 96–112.

[22] Hamed Faroqi and Mahmoud Mesbah. 2021. Inferring trip purpose by clustering sequences of smart card records. *Transportation Research Part C: Emerging Technologies* 127 (2021), 103131.

[23] Gregory Ference, Mao Ye, and Wang-Chien Lee. 2013. Location recommendation for out-of-town users in location-based social networks. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. 721–726.

[24] Huiji Gao, Jiliang Tang, Xia Hu, and Huan Liu. 2015. Content-aware point of interest recommendation on location-based social networks. In *Twenty-ninth AAAI conference on artificial intelligence*.
[25] Qinggang Gao, Joseph Molloy, and Kay W Axhausen. 2021. Trip purpose imputation using GPS trajectories with machine learning. *ISPRS International Journal of Geo-Information* 10, 11 (2021), 775.
[26] Li Gong, Xi Liu, Lun Wu, and Yu Liu. 2016. Inferring trip purposes and uncovering travel patterns from taxi trajectory data. *Cartography and Geographic Information Science* 43, 2 (2016), 103–114.
[27] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. 2008. Understanding individual human mobility patterns. *nature* 453, 7196 (2008), 779–782.
[28] Qing Guo, Zhu Sun, Jie Zhang, and Yin-Leng Theng. 2020. An Attentional Recurrent Neural Network for Personalized Next Location Recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 83–90.
[29] Melissa A Hardy and Alan Bryman. 2004. Handbook of data analysis. (2004).
[30] Samiul Hasan and Satish V Ukkusuri. 2014. Urban activity pattern classification using topic models from online geo-location data. *Transportation Research Part C: Emerging Technologies* 44 (2014), 363–381.
[31] Haibo He and Edwardo A Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering* 21, 9 (2009), 1263–1284.
[32] Thomas Hofmann. 1999. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 289–296.
[33] Bo Hu, Mohsen Jamali, and Martin Ester. 2013. Spatio-temporal topic modeling in mobile social media for location recommendation. In *2013 IEEE 13th International Conference on Data Mining*. IEEE, 1073–1078.
[34] Tâm Huynh, Mario Fritz, and Bernt Schiele. 2008. Discovery of activity patterns using topic models. In *Proceedings of the 10th international conference on Ubiquitous computing*. 10–19.
[35] Amir Khatibi, Fabiano Belem, Ana P Silva, Dennis Shasha, Marcos A Goncalves, et al. 2018. Improving tourism prediction models using climate and social media data: A fine-grained approach. In *Twelfth International AAAI Conference on Web and Social Media*.
[36] Zae Myung Kim, Young-Seob Jeong, Jonghwan Hyeon, Hyungrai Oh, and Ho-Jin Choi. 2016. Classifying travel-related intents in textual data. *Int. J. of Computing, Communications & Instrumentation Engg* 3 (2016), 96–101.
[37] Serhan Kotiloglu, Theodoros Lappas, Konstantinos Pelechrinis, and PP Repoussis. 2017. Personalized multi-period tour recommendations. *Tourism Management* 62 (2017), 76–88.
[38] Sotiris B Kotsiantis, Ioannis Zaharakis, P Pintelas, et al. 2007. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering* 160, 1 (2007), 3–24.
[39] Sang Gu Lee and Mark Hickman. 2014. Trip purpose inference using automated fare collection data. *Public Transport* 6 (2014), 1–20.
[40] Antônio HG Leite, Fabrıcio Benevenuto, and Mirella M Moro. 2013. TripTag: Ferramenta de planejamento de viagens baseada em experiências de usuários de redes sociais. In *28 TH BRAZILIAN SYMPOSIUM ON DATABASES*. 37.
[41] John M Levy. 2016. *Contemporary urban planning*. Taylor & Francis.
[42] Aoyong Li and Kay W Axhausen. 2018. Trip purpose imputation for taxi data. In *18th Swiss Transport Research Conference (STRC 2018)*. STRC.
[43] Huayu Li, Yong Ge, Richang Hong, and Hengshu Zhu. 2016. Point-of-interest recommendations: Learning potential check-ins from friends. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 975–984.
[44] Wenhao Li, Yanjie Ji, Xianqi Cao, and Xinyi Qi. 2020. Trip purpose identification of docked bike-sharing from IC card data using a continuous hidden Markov model. *IEEE Access* 8 (2020), 189598–189613.
[45] Zhishuai Li, Gang Xiong, Zebing Wei, Yu Zhang, Meng Zheng, Xiaoli Liu, Sasu Tarkoma, Min Huang, Yisheng Lv, and Chuheng Wu. 2021. Trip purposes mining from mobile signaling data. *IEEE Transactions on Intelligent Transportation Systems* 23, 8 (2021), 13190–13202.
[46] Ziyue Li, Hao Yan, Chen Zhang, and Fugee Tsung. 2022. Individualized passenger travel pattern multi-clustering based on graph regularized tensor latent dirichlet allocation. *Data Mining and Knowledge Discovery* 36, 4 (2022), 1247–1278.
[47] Defu Lian, Cong Zhao, Xing Xie, Guangzhong Sun, Enhong Chen, and Yong Rui. 2014. GeoMF: joint geographical modeling and matrix factorization for point-of-interest recommendation. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 831–840.
[48] Chengwu Liao, Chao Chen, Suiming Guo, Zhu Wang, Yaxiao Liu, Ke Xu, and Daqing Zhang. 2022. Wheels know why you travel: Predicting trip purpose via a dual-attention graph embedding network. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 1 (2022), 1–22.
[49] Youfang Lin, Huaiyu Wan, Rui Jiang, Zhihao Wu, and Xuguang Jia. 2014. Inferring the travel purposes of passenger groups for better understanding of passengers. *IEEE Transactions on Intelligent Transportation Systems* 16, 1 (2014), 235–243.

[50] Yiding Liu, Tuan-Anh Nguyen Pham, Gao Cong, and Quan Yuan. 2017. An experimental evaluation of point-of-interest recommendation in location-based social networks. *Proceedings of the VLDB Endowment* 10, 10 (2017), 1010–1021.

[51] R Logesh, V Subramaniyaswamy, V Vijayakumar, and Xiong Li. 2019. Efficient user profiling based intelligent travel recommender system for individual and group of users. *Mobile Networks and Applications* 24 (2019), 1018–1033.

[52] Jie Lu, Dianshuang Wu, Mingsong Mao, Wei Wang, and Guangquan Zhang. 2015. Recommender system application developments: a survey. *Decision Support Systems* 74 (2015), 12–32.

[53] Yijing Lu and Lei Zhang. 2015. Imputing trip purposes for long-distance travel. *Transportation* 42 (2015), 581–595.

[54] Washington Luiz, Felipe Viegas, Rafael Alencar, Fernando Mourão, Thiago Salles, Dárlinton Carvalho, Marcos Andre Gonçalves, and Leonardo Rocha. 2018. A Feature-Oriented Sentiment Rating for Mobile App Reviews. In *Proceedings of the 2018 World Wide Web Conference* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1909–1918. https://doi.org/10.1145/3178876.3186168

[55] Suxing Lyu, Tianyang Han, Peiran Li, Xingyu Luo, and Takahiko Kusakabe. 2022. A Dual-Flow Attentive Network With Feature Crossing for Chained Trip Purpose Inference. *IEEE Transactions on Intelligent Transportation Systems* 24, 1 (2022), 631–644.

[56] Suxing Lyu, Tianyang Han, Yuuki Nishiyama, Kaoru Sezaki, and Takahiko Kusakabe. 2022. A plug-in memory network for trip purpose classification. In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*. 1–12.

[57] Chuishi Meng, Yu Cui, Qing He, Lu Su, and Jing Gao. 2017. Travel purpose inference with GPS trajectories, POIs, and geo-tagged social media data. In *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, 1319–1324.

[58] Chuishi Meng, Yu Cui, Qing He, Lu Su, and Jing Gao. 2019. Towards the inference of travel purpose with heterogeneous urban data. *IEEE Transactions on Big Data* 8, 1 (2019), 166–177.

[59] Lara Montini, Nadine Rieser-Schüssler, Andreas Horni, and Kay W Axhausen. 2014. Trip purpose identification from GPS tracks. *Transportation Research Record* 2405, 1 (2014), 16–23.

[60] Attila M Nagy and Vilmos Simon. 2018. Survey on traffic prediction in smart cities. *Pervasive and Mobile Computing* 50 (2018), 148–163.

[61] Minh Hieu Nguyen, Jimmy Armoogum, Jean-Loup Madre, and Cédric Garcia. 2020. Reviewing trip purpose imputation in GPS-based travel surveys. *Journal of Traffic and Transportation Engineering (English Edition)* 7, 4 (2020), 395–412.

[62] Kyo-Joong Oh, Zaemyung Kim, Hyungrai Oh, Chae-Gyun Lim, and Gahgene Gweon. 2016. Travel intention-based attraction network for recommending travel destinations. In *2016 International Conference on Big Data and Smart Computing (BigComp)*. IEEE, 277–280.

[63] Bing Pan. 2010. Online travel surveys and response patterns. *Journal of Travel Research* 49, 1 (2010), 121–135.

[64] Jianpei Qian, Chunfu Shao, Chunjiao Dong, and Shichen Huang. 2021. Inferring Inter-City Trip Purpose From the Perspective of the Group. *IEEE Access* 9 (2021), 6388–6401.

[65] Ettore Recchi and Katharina Tittel. 2023. The Empirical Study of Human Mobility: Potentials and Pitfalls of Using Traditional and Digital Data. In *Handbook of Computational Social Science for Policy*. Springer, 437–464.

[66] Li Shen and Peter R Stopher. 2013. A process for trip purpose imputation from Global Positioning System data. *Transportation Research Part C: Emerging Technologies* 36 (2013), 261–267.

[67] Thiago H Silva, Pedro OS Vaz de Melo, Jussara M Almeida, Juliana Salles, and Antonio AF Loureiro. 2013. A comparison of foursquare and instagram to the study of city dynamics and urban social behavior. In *Proceedings of the 2nd ACM SIGKDD international workshop on urban computing*. 1–8.

[68] Elton F de S Soares, Kate Revoredo, Fernanda Baião, Carlos A de MS Quintella, and Carlos Alberto V Campos. 2019. A combined solution for real-time travel mode detection and trip purpose prediction. *IEEE Transactions on Intelligent Transportation Systems* 20, 12 (2019), 4655–4664.

[69] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. 2010. Limits of predictability in human mobility. *Science* 327, 5968 (2010), 1018–1021.

[70] Peter Stopher and Li Shen. 2011. In-depth comparison of global positioning system and diary records. *Transportation Research Record* 2246, 1 (2011), 32–37.

[71] Haodong Sun, Yanyan Chen, Yang Wang, and Xiaoming Liu. 2021. Trip purpose inference for tourists by machine learning approaches based on mobile signaling data. *Journal of Ambient Intelligence and Humanized Computing* (2021), 1–15.

[72] Kim Carlotta Von Schönfeld and Luca Bertolini. 2017. Urban streets: Epitomes of planning challenges and opportunities at the interface of public space and mobility. *Cities* 68 (2017), 48–55.

[73] Hao Wang, Manolis Terrovitis, and Nikos Mamoulis. 2013. Location recommendation in location-based social networks using user check-in data. In *Proceedings of the 21st ACM SIGSPATIAL international conference on advances in geographic information systems*. 374–383.

[74] Jingjing Wang, Xi Chen, Zhihong Chen, and Lizeng Mao. 2016. Cluster algorithm based on LDA model for public transport passengers' trip purpose identification in specific area. In *2016 IEEE International Conference on Intelligent*

Transportation Engineering (ICITE). IEEE, 186–192.
[75] Jean Wolf, Randall Guensler, and William Bachman. 2001. Elimination of the travel diary: Experiment to derive trip purpose from global positioning system travel data. *Transportation Research Record* 1768, 1 (2001), 125–134.
[76] Guangnian Xiao, Zhicai Juan, and Chunqin Zhang. 2016. Detecting trip purposes from smartphone-based travel surveys with artificial neural networks and particle swarm optimization. *Transportation Research Part C: Emerging Technologies* 71 (2016), 447–463.
[77] Mao Ye, Peifeng Yin, Wang-Chien Lee, and Dik-Lun Lee. 2011. Exploiting geographical influence for collaborative point-of-interest recommendation. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. 325–334.
[78] Anthony GO Yeh. 1999. Urban planning and GIS. *Geographical information systems* 2, 877-888 (1999), 1.
[79] Haochao Ying, Jian Wu, Guandong Xu, Yanchi Liu, Tingting Liang, Xiao Zhang, and Hui Xiong. 2019. Time-aware metric embedding with asymmetric projection for successive POI recommendation. *World Wide Web* 22, 5 (2019), 2209–2224.
[80] Kyung-Hyan Yoo, Marianna Sigala, and Ulrike Gretzel. 2016. Exploring TripAdvisor. In *Open tourism*. Springer, 239–255.
[81] Haitao Yuan and Guoliang Li. 2021. A survey of traffic prediction: from spatio-temporal data to intelligent transportation. *Data Science and Engineering* 6 (2021), 63–85.
[82] Chenyi Zhang, Hongwei Liang, and Ke Wang. 2016. Trip recommendation meets real-world constraints: POI availability, diversity, and traveling time uncertainty. *ACM Transactions on Information Systems (TOIS)* 35, 1 (2016), 1–28.
[83] Chenyi Zhang, Hongwei Liang, Ke Wang, and Jianling Sun. 2015. Personalized trip recommendation with poi availability and uncertain traveling time. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. 911–920.
[84] Jia-Dong Zhang and Chi-Yin Chow. 2015. Geosoca: Exploiting geographical, social and categorical correlations for point-of-interest recommendations. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*. 443–452.
[85] Jia-Dong Zhang, Chi-Yin Chow, and Yanhua Li. 2014. Lore: Exploiting sequential influence for location recommendations. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 103–112.
[86] De Zhao, Wei Zhou, Wei Wang, and Xuedong Hua. [n. d.]. Trip purpose prediction using travel survey data with POI information via gradient boosting decision trees. *IET Intelligent Transport Systems* ([n. d.]).
[87] Shenglin Zhao, Irwin King, and Michael R Lyu. 2016. A survey of point-of-interest recommendation in location-based social networks. *arXiv preprint arXiv:1607.00647* (2016).
[88] Xueliang Zhao, Zhishuai Li, Yu Zhang, and Yisheng Lv. 2020. Discover Trip Purposes From Cellular Network Data With Topic Modeling. *IEEE Intelligent Transportation Systems Magazine* 14, 4 (2020), 37–46.
[89] Xiaoyao Zheng, Yonglong Luo, Liping Sun, Ji Zhang, and Fulong Chen. 2018. A tourism destination recommender system using users' sentiment and temporal dynamics. *Journal of Intelligent Information Systems* 51, 3 (2018), 557–578.
[90] Zack Zhu, Ulf Blanke, and Gerhard Tröster. 2014. Inferring travel purpose from crowd-augmented human mobility data. In *Proceedings of the first international conference on IoT in urban space*. 44–49.