Taxi Destination Prediction based on Points of Interest

Eugenia Clarissa Anjani (03749432)

July 18, 2022

Abstract

Predicting the next destination of taxi journeys is a highly-researched topic in human mobility. The analysis of mobility flows offers insights into the operations of public transport systems to help governments improve urban planning by predicting and reducing traffic. For taxi companies, they can optimize the effectiveness of internal dispatching systems to provide their customers better services by reducing waiting times, saving on fuel, and allocating drivers better. This paper will leverage the ECML/PKDD 15 Taxi Trajectory Prediction dataset which consists of initial partial trajectories of taxi rides in Porto, Portugal. The dataset will be enriched with points of interest (POI) data from Location-Based Social Networks (LBSNs) like Foursquare that can provide the number and type of activities in an area based on predefined categories. This study compares various supervised and unsupervised learning model approaches and how these methods work with and without point of interests data integrated to determine whether or not adding data on hotspots is valuable in predicting the next destinations of taxis. This paper proposes a methodology which extracts movement patterns of taxi trips in Porto based on the trajectory data, while considering human movement patterns that are represented by the POIs. Trips and POIs are clustered using K-Means clustering, each point of the trip is then assigned to the cluster of the point of interest. The contribution of this paper is to determine the value of integrating taxi mobility data with data on POIs.

1 Introduction

Taxi trajectories reflect human mobility over today's modern network of urban roads [Liu et al., 2012]. A significant number of people consisting of residents, tourists and other road users are travelling by taxis or a ride-sharing vehicle in their daily movements around the world due to the conveniences of taxis being a convenient door-to-door and on-demand transport [Zhang et al., 2019]. That is, a person who wants to ride a taxi or ride-sharing vehicle can in ideal cases request for a pickup at any location in a city, and at any time of the day. But the benefits of this location and time flexibility can only be reaped if taxi providers can predict next destinations in different locations to then better manage their electronic dispatching systems.

Predicting the next destination in a taxi journey is a well-studied problem in human mobility that could lend itself to several other real-world scenarios from optimizing the efficiency of electronic dispatching systems to reducing traffic congestion [Rossi et al., 2020]. The improvement of electronic dispatching of taxis not only betters the mobilization of customers through the city in areas with high taxi demand, but it also contributes to optimizing travel patterns and reducing traffic pollution [Tang et al., 2021].

The real-time estimation of destination and travel time for taxis serve great importance for existing electronic dispatch systems, but taxi dispatch planning and coordination is a challenging field given the traffic dynamics and intricacies of human behavior [Lam et al., 2015] [Zhang et al., 2019]. Nevertheless, understanding the behavior of moving objects like humans and cars, can help urban planners find hot spots. With further analysis of human movement patterns, it is possible to predict the trajectory movement and make some suggestions for public transportation routes [da Silva and Fileto, 2015]. Trajectories are valuable inputs for methods that investigate moving object behaviors while understanding what tasks these objects are performing in a certain geographical space like a POI [da Silva and Fileto, 2015]. Discovering areas of interest and behavior of moving objects, both humans and vehicles can help find areas with high energy consumption, thus identifying areas with high pollution levels [Zheng et al., 2013]. For businesses, understanding human mobility could be of interest

to decide on ideal business locations and advertisement spots. In the realm of security, trajectory data mining can also identify places of high threats of public security by examining outlier and anomalies in objects movements [da Silva and Fileto, 2015]. The study of human mobility patterns is very important as it even extends to various other aspects including the control of epidemic spreading, tourism management, transportation prediction and planning, emergency management, and urban planning [Xu et al., 2017]. However, the investigation of travel trajectories like taxi mobility alongside other data sources on human mobility has not been sufficiently exploited, even though can provide even more supporting evidence for various use cases [da Silva and Fileto, 2015] [Smith et al., 2017].

To analyze both human and vehicle mobility, it would be sensible to adopting a data-driven and intelligent approach for analyzing travel demand. Such an approach requires an examination of the predictability of people's travel behavior based on large-scale trajectory data [Tang et al., 2021]. For transportation companies like taxi providers and location-based services, trajectory data mining can provide more personalized navigation services, recommending sightseeing POIs with far-reaching benefits on better transportation management and control [Rossi et al., 2020]. An even more effective management of travel demand is even more important nowadays given the marked increase in the density of urban populations in recent years that put high pressure on urban transportation systems [Ahmed et al., 2008].

In today's data-rich world, trajectory data are collected through sensors, actuators, and mobile devices, that are then analyzed based on network infrastructures [Rossi et al., 2020]. The widespread adoption of positioning services such as Global Position Systems (GPS) and Radio Frequency Identification (RFID) from mobile device sensors has generated a large amount of trajectory data and tracking positions of moving objects such as humans and vehicles [Ibrahim and Shafiq, 2018] [Tang et al., 2021]. Each taxi is supervised by electronic equipment (e.g. GPS device) and network system (e.g. cab reservation system through Uber in Europe, DIDI in China) [Zhou et al., 2016]. These services and devices extract a trajectory, a sample of timestamped location points. The trajectory points and tracking positions can then be further studied to understand the behavior of the moving objects [Ibrahim and Shafiq, 2018].

Each location point in a trajectory serves as a combination of a space and time instance [Ibrahim and Shafiq, 2018]. The GPS trajectory data provides multifaceted information on the position, speed, time, occupancy status of the taxi cab [Zhang et al., 2019]. In addition to GPS data, other sources of information include those generated from mobile phones and Location-Based Social Networks (LBSNs) that can be leveraged to examine human mobility behavior and coordinate city traffic [Rossi et al., 2020]. Common LBSNs include Foursquare that provide the number and categories of activities in a specific area (e.g. ice cream, art museum etc.) that provides insights on the number of people that assemble in that target area or hotspot. The information from such LBSNs can be used to predict better the taxi destinations as it has been used similarly by others (See [Rossi et al., 2020], cite[Tang et al., 2021], [Liu et al., 2012]). Another LBSN could involve geotagged social media posts. For instance, in another Kaggle competition, [Smith et al., 2017] used not only taxi mobility data, but also geotagged tweets from Twitter activity to classify regions of the city into land-use profiles to predict taxi pickup locations based on time, previous pickup counts, and tweet counts.

Taxi destination prediction aims to forecast the destination of the current trip based on the known information of the current trip [Tang et al., 2021]. The accumulation of this trajectory data allows opportunities for investigating destination prediction to generate highly accurate predictions of next taxi destinations. Alongside LBSN data, the collected traffic dynamics and GPS trajectories data represent a valuable source of information to uncover movement patterns and attraction areas [Ibrahim and Shafiq, 2018]. From mining the trajectory data, it would then be possible to bring to light some insights in the area of routes and destination prediction for the purposes of urban and traffic planning.

This paper revisits a former Kaggle competition related to the taxi destination prediction challenge that was organized by the 2015 ECML/PKDD conference [Kag, a]. The task of the challenge was to find the destination of taxi trips based on initial partial trajectories. The paper examines the provided data set with approximately 1.7 million taxi journeys from 442 taxis cruising the city of Porto, Portugal, for a complete year from 2013-07-01 to 2014-06-20. The goal of this paper is to improve the taxi electronic dispatch by estimating the final destination of the total trip of ongoing taxi rides based on their initial partial trajectories while taking into account patterns of human mobility that are represented by POI. As a result, this paper contributes to providing a data-driven solution to the former Kaggle challenge

with an analysis of human movement patterns. This paper examines various machine learning model approaches and how these methods perform with and without the POI data based on some common metrics and the Kaggle leaderboard. The object is to then determine whether or not adding data on hotspots is valuable in predicting the next destinations of taxis.

The remainder of this paper is organized as follows. Section 2 summarizes the related literature of taxi destination prediction with a view on passenger and taxi driver patterns. In Section 3, a comprehensive analysis of the taxi mobility and POI data is provided. Section 4 explains the methodology with further insights into the selected approach, technical setup, data preprocessing and model prediction. Section 5 informs about the results that were obtained on Kaggle based on the leaderboard. In the following section 6, the alternative approaches are discussed. Finally, the paper ends with section 5 that will present the limitations and recommendations on future research steps.

2 Literature Review

The ongoing development of urban computing and object movement studies, as well as abundant trajectory data provides a rich body of resources to examine intra-urban human mobility patterns. [Liu et al., 2012] [Kang and Qin, 2016]. It is difficult to understand mobility patterns because it is intertwined with complex human behavior and different taxi drivers' cognition [Zhang et al., 2019]. Nevertheless, the following two sections examines the existing literature related to two key stakeholders in the area of taxi driving trajectories: (1) taxi passengers and (2) taxi drivers.

2.1 Passengers' mobility patterns

The mobility patterns of passengers has been a well-studied problem in geography, transportation, and urban studies. Prior studies on taxi travel patterns have focused on evaluating passenger travel patterns from passenger trips. Tracing back all the way to the 1970s, Hägerstrand demonstrated his initial space-time framework to model the travel behaviors of individuals and discuss themes in human geography [Hägerstrand, 1970]. Since then, the study of travel activities became a widely research problem [Kang and Qin, 2016]. Further compounded by the development of GPS and RFID data from mobile device sensors has allowed taxicabs to have digital traces that have contributed significantly to the understanding of human movements [Ibrahim and Shafiq, 2018] [Tang et al., 2021] [da Silva and Fileto, 2015]. Large geospatial data, which are extracted from sources such as mobile phone records, social media check-ins, and taxi trajectories, provide abundant locations to model movement of people around cities [Smith et al., 2017]. The available literature views patterns in taxicab passengers as representative of human movement patterns [Zhou et al., 2016]. This can also be seen from the analysis of the dataset in the third section of this paper.

The focus of current literature seem to lie in the spatio-temporal distributions of trip origins and destinations and to what extent that they taxi pick-ups and drop-offs concentrate in these spatial and temporal dimensions [Kang and Qin, 2016]. Taxi GPS traces can also be used to plan for other transportation modes like what was done in for planning night-bus routes in [Chen et al., 2014]. In this study, the main approach was to use the taxi GPS traces to select best bus routes which expects the maximal number of passengers under certain conditions such as bus operation time constraints and traffic conditions. In order to understand further the observed patterns of human movement, some studies exist on land use. Much of the current literature examines the interdependence between urban land use and travel patterns with a focus on the estimation of traffic or trip lengths from land use [Chen et al., 2009]. Other literature also investigate the different time variations of taxi origin density given some land use conditions at the original destination. In [Liu et al., 2012], the authors investigated the temporal variations of both pick-ups and drop-offs, and how different land use features, types and intensity play different roles at different times a day in a trip generation. In relation to the motivation of this paper that assumes the behaviors of passengers are predictable, a significant amount of literature also exists. The literature in the area of the regularity of activity venue and travel route selections describes the high predictability of taxi trip origins and destinations in cities. In the area of urban hotspot detection, much literature has been done in this area to reflect residents travel characteristics and the operational status of urban traffic to determine taxi pick up and drop off [Yu et al., 2021] [Zhou et al., 2018]. In the realm of urban traffic, [Gallotti et al., 2015] shares that speed depends on the duration of the trip, with larger speed for longer trips. This study also further explains some

patterns on length of trips, that people may prefer shorter taxi rides during periods of high traffic. In another paper by [Peng et al., 2012] describe the high predictability of identifying taxi trip origins and destinations in cities based on their daily commute that is further supported by probabilistic distributions. These studies shed bright light on the repetitious daily routines and the highly limited activity budgets of city dwellers. These studies provide insights on the repetitive daily routines of residents, which can also be observed from the data analysis in section 3.1.2 .

2.2 Taxi drivers' mobility patterns

The behaviors of passengers and drivers are two critical factors that impact the operations of taxis. In pursuit of striving for operational efficiencies, various studies lead towards the drivers strategy for gaining profit. Such operational behaviors of taxi drivers are reflected in their actions to hunt, wait, and shift their driving schedules [Li et al., 2011] [Yuan et al., 2011].

From the perspective of where to find the next passenger (i.e. hunting behavior), existing literature has pointed towards searching strategies based on hotspot detection [Yu et al., 2021] [Zhao et al., 2015] [Zhou et al., 2018]. From these studies, hotspots for taxi boarding are pooled in various areas. In the study of [Zhou et al., 2018], the authors show that passenger pick-up hotspots for taxis in Nanjing were mainly concentrated in specific districts, which can allow better recommendations for current passenger pickup hotspots for taxi drivers, thus increasing the passenger-carrying rate of taxis, increasing income, and at the same time reducing vehicle pollution. In addition to hotspots, hunting behavior should also consider time (as a proxy of curising distance) and time (as an indicator of traffic condition) [Yuan et al., 2012].

Considering the next strategy of where to, if not hunt, wait for potential passengers? And how long to wait (i.e. waiting behavior) [Kang and Qin, 2016]. This behavior is strongly connected to the hunting behavior where in most circumstances the cab drivers would choose to wait at, or near by to a previous drop-off location if the current location is a hotspot with a high potential for passenger pickups or just that it is too costly to hunt for the next passenger [Li et al., 2011].

In the realm of when and where to interrupt taxi working behaviors, such issues might involve refueling, resting, shifting, and stopping behaviors [Kang and Qin, 2016]. Such daily decisions of drivers might depends on urban conditions such as the location of gas stations and parking lots [Kang and Qin, 2016].

Other existing studies point towards drivers behaviors related to their cognition level and the external environment. These factors have been said to impacts drivers' operational behaviors [Li et al., 2011]. To meet the demand of city travels and earn more income, taxi drivers need to acquire self-organized habitual operation behaviors in space and time based on their operating experience and cognition map of the urban built environment [Kang and Qin, 2016]. Some studies also examine drivers' misbehavior. In [Zhou et al., 2016], the authors used taxi's big data in real-time to detect taxi drivers' misbehavior where passengers are detoured and overcharged. Taxi driver fraud is apparently a significant problem because it is difficult problem to obtain solid evidence to expose the fraud and results in further problems in managing the taxi monitoring system [Zhou et al., 2018].

In this perspective, it would be desirable to help drivers decide which of two potential rides will be more profitable. A related to study with respect to fare data in Bangkok [Phiboonbanakit and Horanont, 2016] shows that drivers tend to take customers to destinations that is different from the customer's original request. However it would actually be meaningful for drivers to know that on average the short journey also makes more profit than the long-distance journey when use the same amount of distance range (multiple short trip compare with single long trip) according to the study by [Phiboonbanakit and Horanont, 2016]. Second, the probability of getting customer also depends on the area of interest. The area which are well-known such as department store or city center have high probability to occur taxi trip.

3 Data

3.1 Taxi Mobility Data

Taxi trajectory data is beneficial for the assessment and optimization of transportation infrastructures and policies. There are other taxi trajectory datasets such as Kaggle challenges for New York and San

Francisco, but this paper focuses on Porto, Portugal. The following sections describe the dataset of initial partial trajectories and provide an explanation of the analysis that was performed.

3.1.1 Dataset Description

The dataset is based on a Kaggle competition in the context of the ECML/PKDD 2015 challenge [Kag, a]. The raw CSV file was extracted with a total size of 2 GB and contained approximately 1.7 million records of 442 taxi vehicles running in Porto, for a time period ranging between July 1, 2013 and June 30, 2014. Each record is composed of a taxi trajectory and corresponding external information. The trajectory contains a list of GPS locations every 15 seconds and the last location represents the trip destinations. Each trip tuple consisted of 9 attributes with the trip id as the unique identifier to distinguish among the different trips. The provided external information includes the taxi ID, the departure time of the trip, and the call type. The call type identifies the way passengers called for taxi services. If the trip was dispatched from the central, the call type is denoted by 'A'. The call type is 'B' when a passenger calls taxi services at a specific stand. The call type equals to 'C' when a trip starts randomly on a street. Depending on the call type, the origin call attribute indicates the phone number used to request the taxi, and origin stand attribute indicates the start point of the trip if the start was from a taxi station.

In the set up of the Kaggle competition, the test dataset consists of 320 partial trajectories that were created from five snapshots taken at different timestamps. The testing dataset is divided into two subsets of equal size, a public and private test set. The public dataset was used through the competition to compare models, while the private data set was used only at the end of the competition for the final leaderboard.

3.1.2 Dataset Analysis

In the exploratory data analysis, an analysis on the number of trips for items related to the location, time, and individual route analysis were done.

Pick-up and Drop-off Distributions The first analysis that was done is comparing the pick up and drop off distributions. Figure 1 and 2 show the distributions for the pick-up and drop-off of passengers in the data set. From comparing the two distributions, it looks like that most of the journeys cover a similar trajectory scope.

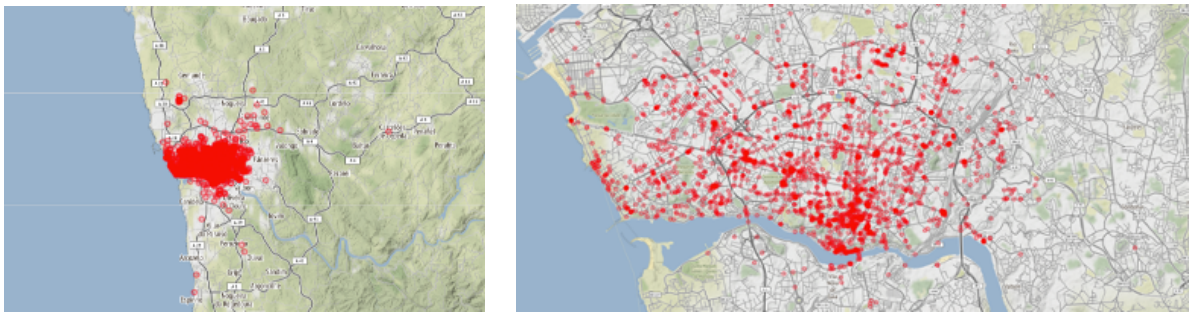

Figure 1: Pick-up distributions in Porto

Number of Trips per Call Type
The highest number of trips was for the trips which were dispatched from a taxi stand with 817K trips. Trips which started from a random street were second highest with 527K trips while trips that started from the central area came last with 364K trips. Such information could provide an incentive for taxi providers to request urban planners or property managers for a taxi stand.

Number of Trips per Month
In terms of the number of trips per month, it can be seen from figure 4 that the highest number of taxi rides occur in May and the second highest in October. When drilling down into the number of trips per month and the days in those months, the highest number of taxi rides occur on the first of January.

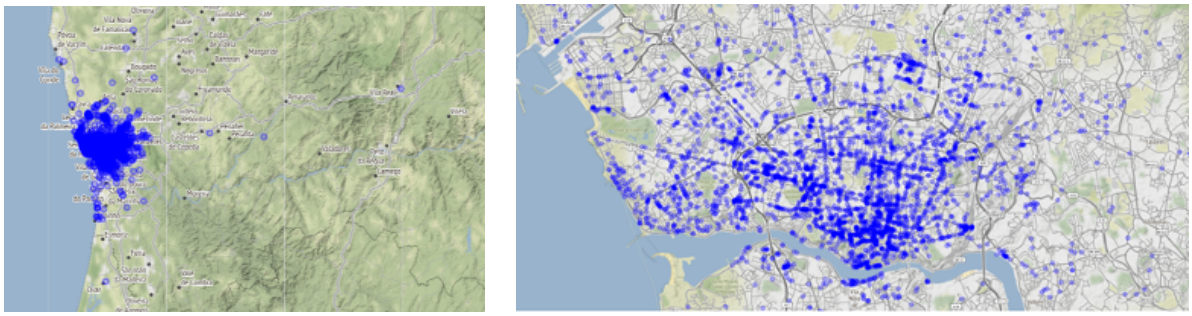

Figure 2: Drop-off distributions in Porto

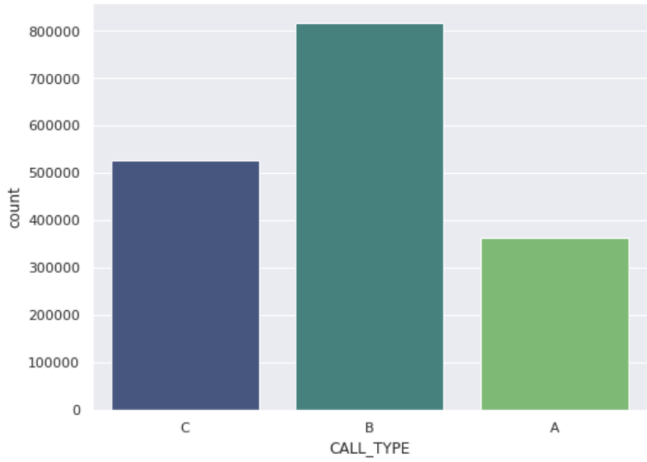

Figure 3: Number of trips per call type

There could be various reasons for the high number of trips. One reason could be Porto's popularity among tourists. For the fourth consecutive year, Portugal was voted as 'Europe's Leading Destination' and the city of Porto was named 'Europe's Leading City Break Destination' at the 2020 World Travel Awards [Por, 2020]. In addition, May and October have lower rainfalls, considering that rainfalls season in Porto is between November until March every year, thus making the weather a pleasant one for tourists to go sightseeing. The tourist season could also play a role in the taxi mobility patterns, because the peak season is not in May and October, but in the summer months of July and August when travel prices are high which thus further incentivizes tourists to visit Porto during the off-season.

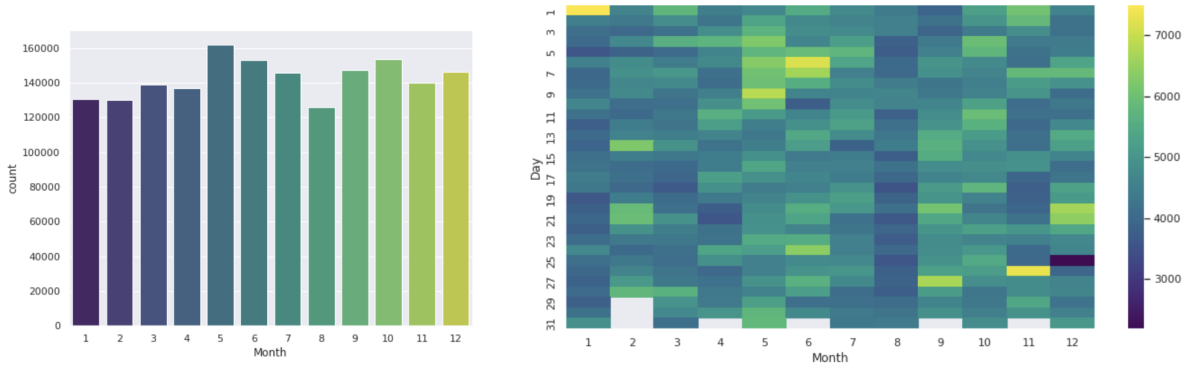

Figure 4: The image on the left shows the number of trips per month, while the image on the right shows the number of trips per month based on the day

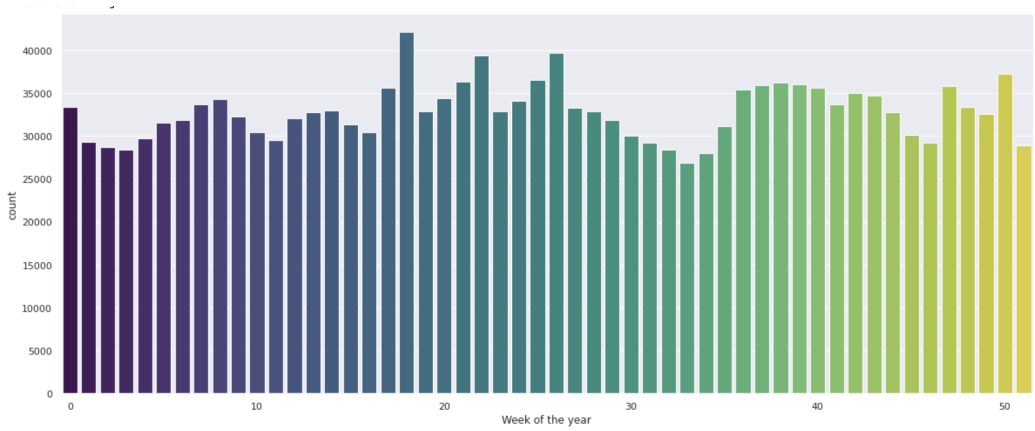

Figure 5:

Number of Trips per Week of Year

The top three number of taxi trips tend to occur in the week 18, 26, and 22 of the year which can be seen in figure 5. These high numbers could be correlated to the festivals and events that are occurring in Porto during the weeks.

Number of trips by Hour of Day

From the graph in figure 6, another angle of city life can be seen by analyzing activities on the weekends and weekdays. On the weekends people will likely be starting the day later compared to weekdays, and city activities can go beyond midnight hours on the weekends.

It can be seen that on the weekdays the highest number of taxi rides occur at around 9:00 AM and then drop in the evening and night hours. The number of trips start to increase at 8:00 AM until it peaks at 9:00 possibly due to rush hour time on the weekdays. It would be reasonable to extrapolate that people tend to be more active in the early mornings due to work and school with the taxi movement decreasing from 10:00 AM for the next hours until the next weekday starts again.

Whereas on the weekends, people tend to be more active in the evening and night hours to enjoy some city attractions. This pattern is further confirmed by right image of figure 6 based on the noticeable pattern that shows taxi rides occurring more early at dusk from Saturday to Sunday. The number of trips was lower at around 1:00 - 2:00 AM but increased significantly until it reached a peak in the early morning hours at approximately 3:00 - 5:00 AM.

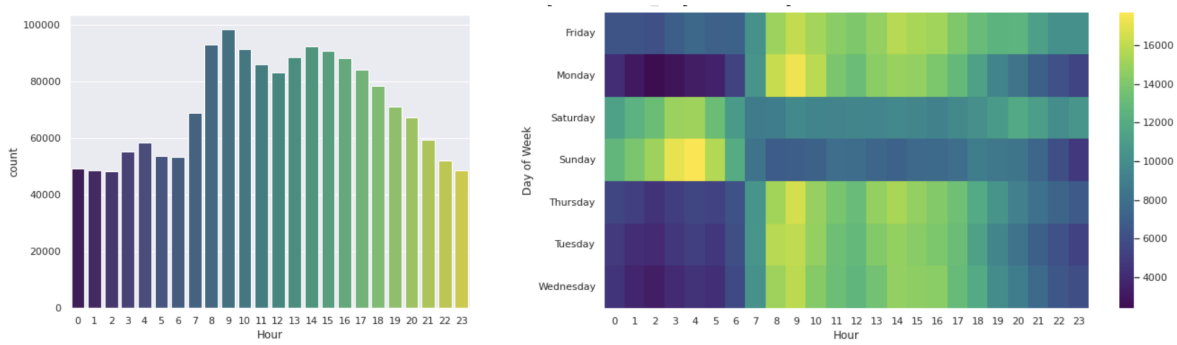

Figure 6: The image on the left shows the number of trips per hour, while the image on the right shows the number of trips per hour based on the day of the week

Trip Duration

From the image, it can be extrapolated that taxi passengers tend to prefer short trips with an average of 7 minutes as shown in figure 7.

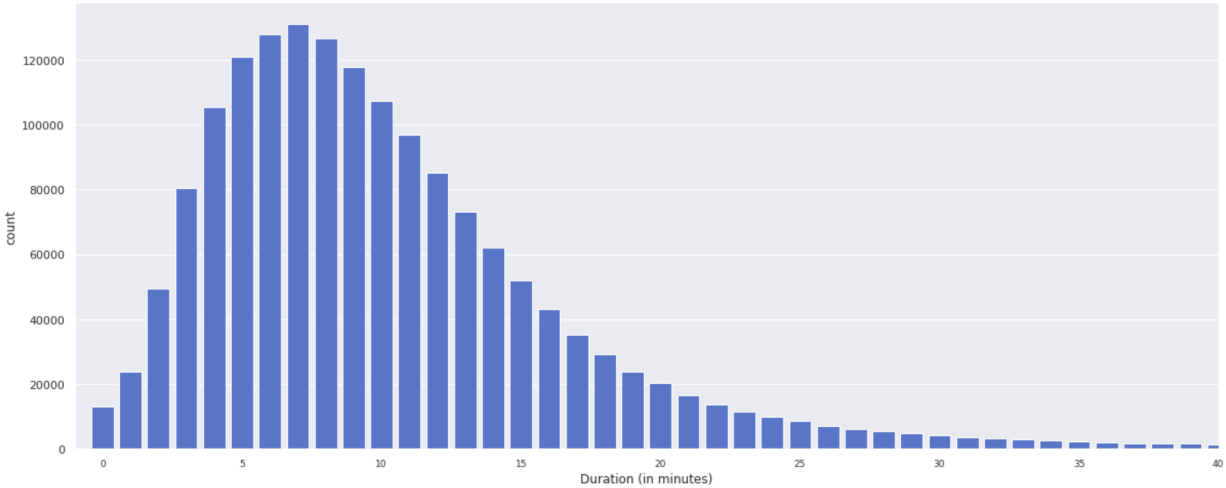

Figure 7: Distribution of trips and their durations

Destination Analysis

The hypothesis of the approach is based on the fact that two trips with similar routes should likely end at the same or nearby destination. For instance, trips to the airport seem to happen very regularly and are quite easy to predict.

In addition, some consistencies in the data that can be observed such as destinations that are frequently visited by the same call ID. There are a number of call IDs that make an exorbitant number of trips which may put the data into question or it may just be that these customers use taxis as their main mode of transportation. The top five frequent call IDs are listed below in table 1.

Call ID	Number of recorded destinations
2002	57,487
63882	6,401
2001	2,488
13168	1,314
6728	1,114
This approach	2.88

Table 1: Top five frequent Call IDs

In viewing the heat map of destinations of various trips, it can be seen that some customers only go to some major locations. In the extreme case of original call ID 2002 in the training data as shown in figure 8, who took 57,487 trips it can be seen that this person took about 19,000 unique destinations but probably this number is smaller if accounting for the fact that there may be only small discrepancies in the coordinates of the destination point.

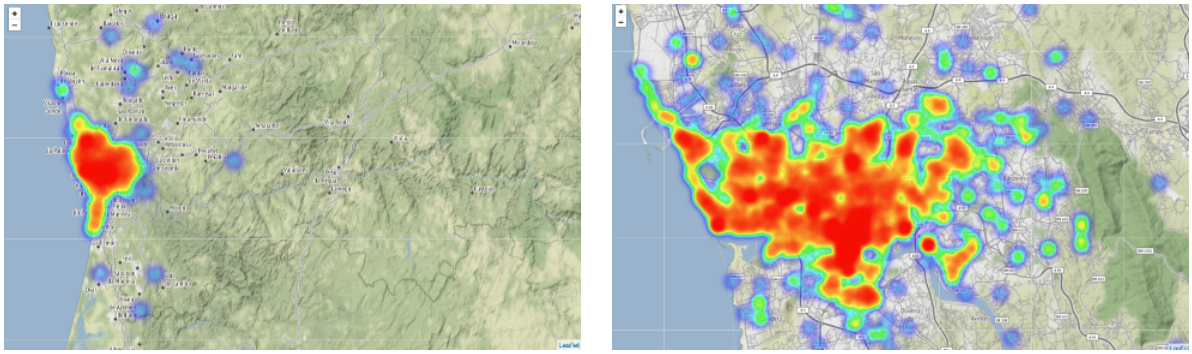

Figure 8: Left image shows the destinations of customer with call ID 2002

In a less extreme case of original call ID 34861 who took 210 trips, heat map shows 15 major destinations, although overall there are 155 trips. When looking at the trips of other call IDs, come patterns of major destinations appear that could represent someone's home, work place, or other personal location.

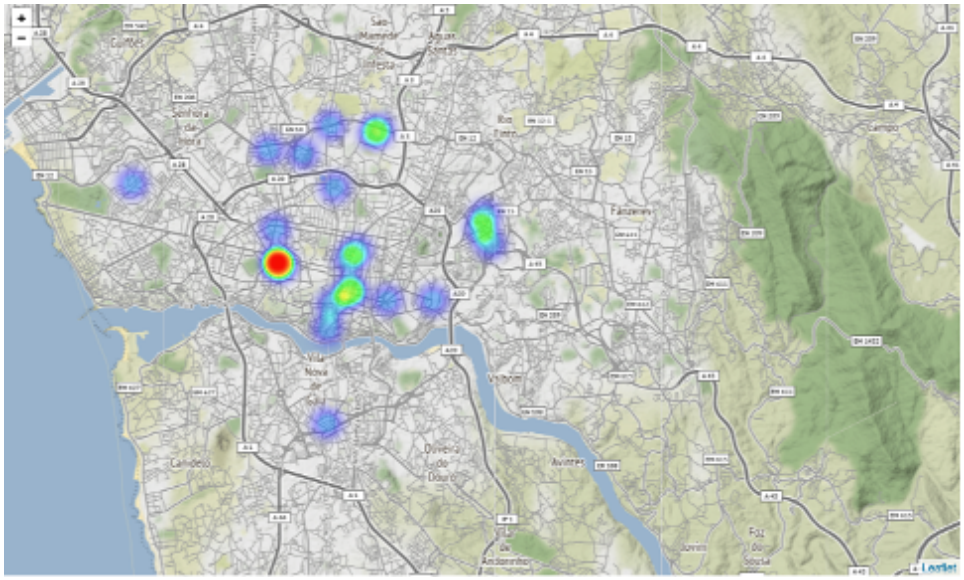

Figure 9: Trips taken by call ID 38461

3.2 Points of Interest Data

Like taxi mobility data, POI data provide insights into key hotspots that residents, tourists, and other individuals go to. For example, knowing hot spots on particular days in modern cities can help in finding areas of high energy consumption or high risks to security. The following sections describe the used POI data and provide an explanation of the analysis that was done.

3.2.1 Dataset Description

Cities like Porto are concentrated areas of human activities which make intra-urban travel a significant part of life for city dwellers. Identification of hotspots plays a role for many applications such city infrastructure, construction, urban transportation planning and management, and many others. In order to provide further information about the activities of individuals, the taxi mobility data in this paper's approach is enriched by data on POIs.

To formalize the definition, a POI is a place of interest with a specific location. POIs are usually museums, beaches, parks, tourist sites, restaurants, and many more. The POI data that was used in this study was obtained from Foursquare's Places API. This API provided various categories to describe a location and activity, including restaurants, beaches, parks, scenic lookouts and so on. This study used version 3 of the API which allowed an extraction of 67 POI. Version 3 of the Places API is the latest version, while previous approaches used version 2 of the API which has already been deprecated for new implementations. Similar approaches that integrated the Porto taxi trajectory dataset used version 2 of the API (See [Rossi et al., 2020], [Tang et al., 2021], [Liu et al., 2012]).

It might be better to merge the model with another POI dataset as only 67 POIs were extracted from version 3 of the Places API. In [Rossi et al., 2020], the authors were able to extract 8928 POIs for Porto with the FourSquare API version 2. There is a possibility to also use a flat file but it was not tested for this seminar study.

3.2.2 Dataset Analysis

In the POIs map, there are points that are located outside of Porto which could be due to the fact that people from Porto make day trips that then bring them out of Porto. Based on the map there

Category	Count
Portugese Restaurant	9
Park	7
Beach	6
Garden	3
Scenic Lookout	3

Table 2: An example table.

are various categories of locations with Portugese restaurants being the highest with 9 locations.

The queried dataset provided 38 different categories of POI. Table 2 shows the top 5 categories that appear more frequently in our dataset of 67 POI. Other categories that only appear once or twice include breweries, barbecue joints, ice cream parlors, monuments, night clubs, and many others.

The results are visualized in an interactive map as shown in figure 10.

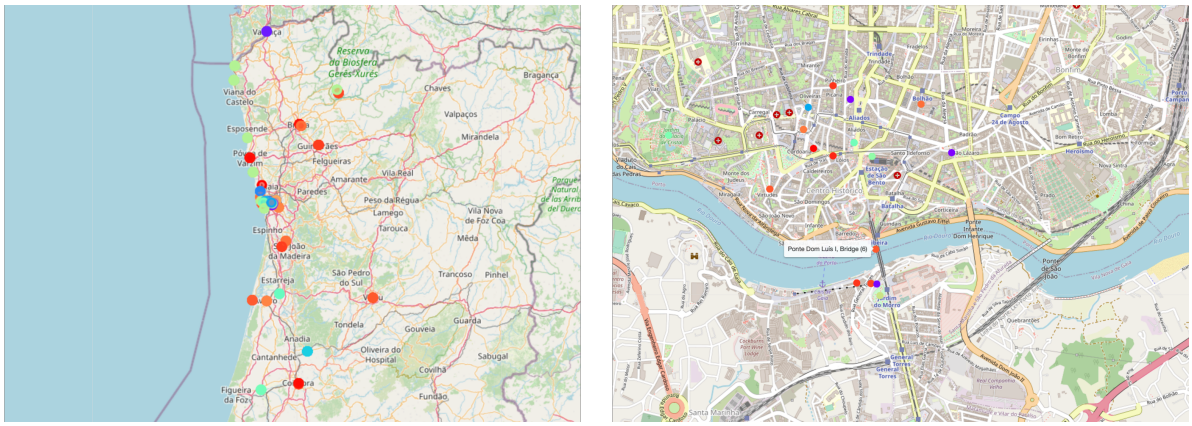

Figure 10: Left image shows a zoomed out image of all the marked POIs and the right image shows a zoomed-in map. Upon hovering on the points on the right image, it is possible to see further details of the location. In the details of the hovered point, a description of the Ponte Dom Luis I bridge is shown with a "Bridge" as the category.

4 Methodology

Understanding patterns of human and taxi mobility in cities has many practical applications for urban management and transportation intelligence. Such an evaluation can provide understanding on the operational status of taxicabs and identify the traffic demand and congestion. The result could then provide data-driven recommendations for operators and passengers for better managing and using taxicabs. The next sections will examine the methodology used to analyze the partial trajectory data of taxis and POI in Porto to study human mobility patterns. The subsections provide information on the general framework, technical setup, data preprocessing, model training and evaluation, as well as model evaluation.

4.1 General Framework of Selected Approach

As part of the general framework, the study compares the result of using the POI data in the machine learning model and without using the POI data. The aim of this approach is to examine whether knowledge of POI locations would result in a considerable improvement in the results. Such an approach of using the POI data in the partial trajectory dataset is not a novel one, but there was no such comparison found to determine if there is any value in adding the POI dataset. To my knowledge, there was no analysis done on the models performance for the comparison of without and with the point of interests data.

The study has a basic approach without the POI data where various regression methods were used including the random forest, gradient boosting, k-Nearest Neighbors, decision tree, and Histogram-based Gradient Boosting were used. Based on a review of the approaches that were chosen in the Kaggle competitions, it was not often the case that a regression was used and the authors of [Rossi et al., 2020]think that they are one of the few that have attempted to frame the challenge as a regression problem. One of the few regression approaches was done in [Rossi et al., 2020], as well as [Smith et al., 2017]. In the work of [Smith et al., 2017], a random forest regression was adopted to predict the taxi pick-up zones in New York City based on the start or end points of trajectories and Twitter data.

Once the POI data were integrated into the basic model, the model uses the K-Means clustering algorithm to cluster the destinations and POI. The destination points were assigned to the clusters of the POIs with the goal of providing a more accurate view of the area. In the next subsections, details for the technical setup, model prediction, and results are provided.

4.2 Technical Setup

Two Google Colab files were used with Python 3.9. The following files have both been shared via the submission process.

- AS-LSCM.ipynb: the main Google Colab notebook which contains all information on the processing of the taxi mobility and FourSquare data
- AS-FourSquare.ipynb: the Google Colab notebook which contains the code to process the Foursquare version 3 API

With regards to the data, the main taxi mobility data is extracted from the Kaggle API that is installed in Google Colab. To authenticate with Kaggle, it is required to create an API token from the individual account which will then trigger the download of kaggle.json, a file containing the API credentials [Kag, b]. Once already authenticated, it is then possible to download the dataset with the following command:

1. !kaggle competitions download -c pkdd-15-predict-taxi-service-trajectory-i

The file for the generated POIs data is in the form of an excel file that was then read into the Google Colab file and turned into a data frame.

4.3 Data Preprocessing

From the extensive data analysis, the dataset was found to show some inconsistencies. It can be observed that there could be some unreliable information. Even from a first glance of the generated map image, it can be seen that there are some exceptional trips that could be erroneous. The Kaggle dataset included a MISSING_DATA column to mark as false when the GPS data is complete, and true whenever one or more locations are missing but it seems that most values of the columns were false even though the records included missing information. The authors of [Lam et al., 2015] noticed that there were missing points in the trajectory by observing the large jumps between the GPS updates based on considering the distances that a taxi would have driven at a speed limit of, for example 160 km/h. Moreover, some trajectories look unusual with coordinates lying outside the expected trajectory shape because perhaps the driver forgot to turn on or off the meter. Some GPS points also lie too close together which might be because of taxis parked with the meter on or perhaps really bad traffic congestion.

With these data issues in mind, the data was preprocessed to extract the features accordingly and serve as input to build the predictive models for destination prediction. The dataset was also cleaned to provide a better picture of the data analysis. The following shows what preprocessing were performed on each of the different fields:

1. **TIMESTAMP**: The format of the timestamp was converted into a datetime object and datetime object. The formatted timestamp was then further broken down into year, month, week of year, day, day of week, weekday, hour, and minute

2. **MISSING_DATA**: The values were changed from False to 0 and True to 1. In addition, any item that had MISSING_DATA was filtered out.

3. **CALL_TYPE**: Values that were numerical were changed to categorical values

4. **ORIGIN_STAND**: If the value is None, the value is changed to 0. If not, the value was changed to 1 or its original value

5. **ORIGIN_CALL**: If the value is None, the value is changed to 0. If not, the value was changed to 1 or its original value

To treat outliers, the origin and destination points that have coordinates three standard deviations below and above the mean were removed.

In addition to the data preprocessing as mentioned above, some additional features were added as follows:

1. **Haversine distance**:
   The Haversine distance between the origin and destination was added as a feature based on the approach of Kaggle challenge participants such as [Lam et al., 2015] [Besse et al., 2018]. Instead of the Euclidean distance, the Haversine distance is used which is the equation that looks for the distance of an arc between two points on longitude and latitude [Kag, a]. The Euclidean distance is not appropriate in this approach because it is used to measure the distance between two points on a flat plane so the results have differences when compared to the Haversine formula. In the approach of [Lam et al., 2015], the authors actually include both the Euclidean and Haversine distance as additional features. The formula of the Haversine distance that was used is as follows from [Gade, 2010], [Kag, a]:

   $$d = 2rarcsin(\sqrt{sin^2(\frac{\varphi_2-\varphi_1}{2}) + cos\varphi_1 \cdot cos\varphi_2 \cdot sin^2(\frac{\lambda_2-\lambda_1}{2})})$$

   where

   (a) φ_1, φ_2 are the latitude of point 1 and latitude of point 2

   (b) λ_1, λ_2 are the longitude of point 1 and longitude of point 2

2. **First and last coordinates consisting of longitude and latitude**:
   The coordinates of the destination and origin were retrieved from the polyline to increase efficiency of the model time. Filtering out only a certain number of points seemed to be a common approach looking at contestants such as [Rossi et al., 2020] and [Kag, d]. Others also used the first-n location points and last-n location points. For instance the approach of [Rossi et al., 2020] considers only the first five and last five trajectory points. Some contestants even included the whole trajectory data instead of using just the origin and final destination. Considering the whole GPS trace further confirms the computation effort due to the huge amount of data to be stored. For the purposes of this paper, only the first and last coordinate points were considered.

The model clustered both the coordinates of the POIs and destinations. In the approach where the POI data were integrated, two clusters were added as features to the training dataset:

1. K-Means cluster assignments that were derived based on the destination points of the taxi mobility data

2. K-Means cluster assignments of the destination points to the clusters from the POIs data

The basic model then included only the call type category, origin call ID, origin stand ID, missing data column, coordinates of the origin, and Haversine distance. In the other model with the POI data, the clusters of the destination and clusters assignments of the destination points to the POI data were included.

4.4 Model Training and Prediction

The next destination problem was treated as a regression problem with continuous destinations as shown in [Rossi et al., 2020]. In the model prediction, only a sample of 5K points was included in the approach as larger samples resulted in unreasonable computational time or crashes in the Google Colab file. In the approach of [Kag, d], a sample of 5K points was used in the model training. The following two subsections explains the approach for the basic approach with and without the POI data.

4.4.1 Basic Approach Without Point of Interests Data

For the basic approach without the POI data, various regression methods were used from the Sklearn library of Python. With the processed data and features, various regression methods were implemented to predict the next taxi destination. Some of these regressions came from the approach of [Kag, d], and other challenge participants. The following regression methods were used:

1. **Gradient Boosting Regression**:
   The regression computes the difference between the current prediction and the known target value, with the difference being called the residual [Skl, b]. The regression can train a weak model that maps features to the residual that is then added to the existing model input [Skl, b]. As a result, it nudges the model towards to correct targets and after some repeated steps, the model prediction is expected to improve. A Gradient Boosting Classification also exists, but the regression is used because the study needs to predict continuous values like the destination's longitude and latitude.

2. **Random Forest Regression**
   The idea behind using the Random Forest regression was from the approach of [Smith et al., 2017] in analyzing the pickups in New York. The nonlinear nature of Random Forest could have some benefit over linear algorithms, to aggregate the results of the decision trees and produce an optimal result. Some approaches online showed an approach of using random forests to predict locations, it was particularly common in a similar Kaggle competition for trip predictions on New York data, see [Rpu,]

3. **K-Nearest Neighbors Regression**
   In the K-Nearest Neighbors (KNN) regression, the association between the independent variables and continuous outcome that is the longitude and latitude were done by averaging the observations in the same neighbourhood [Skl, d]. In the case of this study, given a pick-up point, the regression could output the coordinates of the closest cluster centroid. In the approach of [Yu et al., 2021], the authors proposed a spatial clustering algorithm for detecting urban hotspots based on real-life trajectory data which they claim produced high-quality clustering results.

4. **Decision Tree Regression**
   In bundling the regression model in the form of a tree structure, the decision tree regression allows a dataset to be broken down into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed [Skl, a]. This seemed to be an approach taken by several users in the a similar Kaggle competition on taxi trip prediction on New York data, see [Kag, c] and [Smith et al., 2017]

5. **Histogram-based Gradient Boosting Regression**
   Histogram-based algorithm builds gradient boosting ensembles of trees [Skl, c]. The Histogram-based Gradient Boosting Regression is a common technique for regression predictive modelling due to its ability to perform reasonably across a wide range of datasets. To my knowledge, there was no other competitor that used the Histogram-based Gradient Boosting algorithm on the Porto dataset. It unfortunately can be slow to train the model, especially on large datasets, this can also be seen on the longer computation times in the results section of this paper.

In the implementation, a $MultiOutputRegressor$ was used to support all the above-mentioned regressions that do not natively support multi-target regression. After running all the regressions, a submission file in the form of a CSV is then generated based on the results of the best regressions.

4.4.2 Approach with Point of Interests Data

Supplementing the base approach, the study adds POI data from Foursquare. In the area of spatial distribution, clustering methods have been widely applied to trajectory-based hotspot detection, but the current clustering algorithms for hotspot discovery has some difficulties in meeting requirements of trajectory data and their heterogeneous spatial distribution [Zhao et al., 2015]. In this section, experiments of trajectory clustering are carried out to cluster POIs and taxi mobility data with K-Means clustering because spatial clusters could contribute to improvements in the model.

The model clustered 10 percent of the dataset of destination points which should provide a good enough sample size. Due to significant issues with run time, the dataset for analysis was reduced. The algorithm starts with determining the optimal number of clusters into which the data may be clustered. In this approach, the Elbow Method is used to determine the optimal value of the clusters k.

In the implementation, the K-Means clustering methods from the Sklearn library of Python was used on the destinations of all the training trajectories. Let u be a taxi driver and taxi trajectory $T_u = p_1, p_2, \ldots, p_k$ as an ordered sequence of points based on time that alternates between pick-up and drop-off. Each point of the trajectory is then assigned to the closest centroid c_i where the trajectory T_u is assigned to a new mapped cluster trace $CT_u = c_1, c_1, \ldots, c_k$ where c_i is the identifier of the closest cluster to the point. It is also worth mentioning that there was another Kaggle challenge participant (see [Ibrahim and Shafiq, 2018]) that started to use K-Means clustering but switched to HDBSCAN clustering, an enhancement of the DBSCAN clustering algorithm. The authors actually started with applying K-Means clustering but found that it was not suitable for the analysis because they did not have enough information of the attraction places in Porto which this study compensates for. This study might not have enough data on the POI as well, but it at least has some data, while the authors in [Ibrahim and Shafiq, 2018] did not consider any POI data. The authors also mention the issue with far places of the city that should be considered as noise. Since this study's approach removed the outliers, the magnitude of the issue should be alleviated. The approach of this study could still have some room for improvement as it is a comparatively simple clustering approach compared to other complex clustering methods as mentioned in the use of HDBSCAN clustering in [Ibrahim and Shafiq, 2018] or the Mean-Shift Clustering approach in [Tang et al., 2021].

In this part, the details of how the trajectories are clustered is provided.

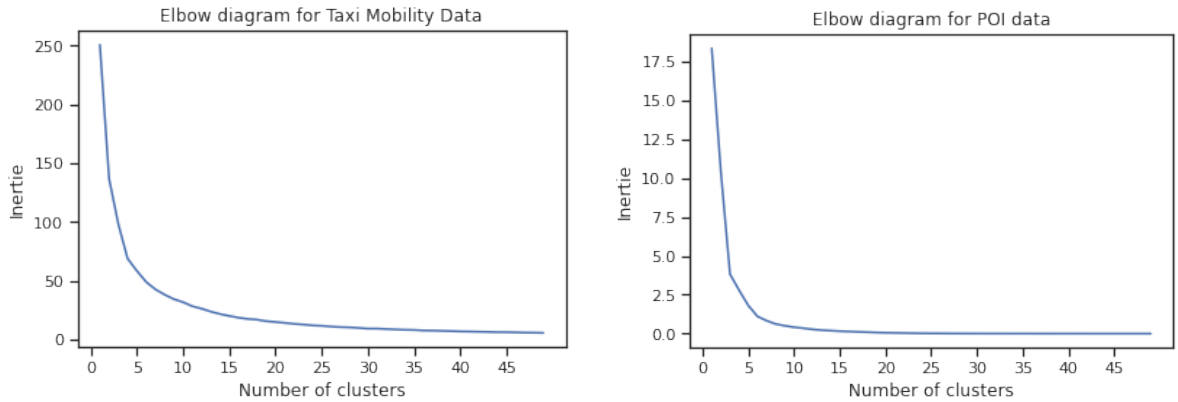

Figure 11: Elbow diagram suggests to have seven clusters for the taxi mobility data as shown on the left diagram and six clusters for the POI data as shown on the right diagram

The number of clusters could be debatable as some studies have used from 45 clusters to several thousand clusters. For instance in [Rossi et al., 2020], 3,392 clusters were used for the Porto dataset. However, according to [Rossi et al., 2020], the model seems insensitive to the number of clusters up to a certain minimal level. It seems that with a high enough number of clusters, the distance between the centroids is very small, and there is no noticeable improvement from adding more clusters as the surface is already well covered. It is still however an open question if the study reached the desired minimal number of clusters. This study relies on the Elbow study which should provide a sound enough cluster.

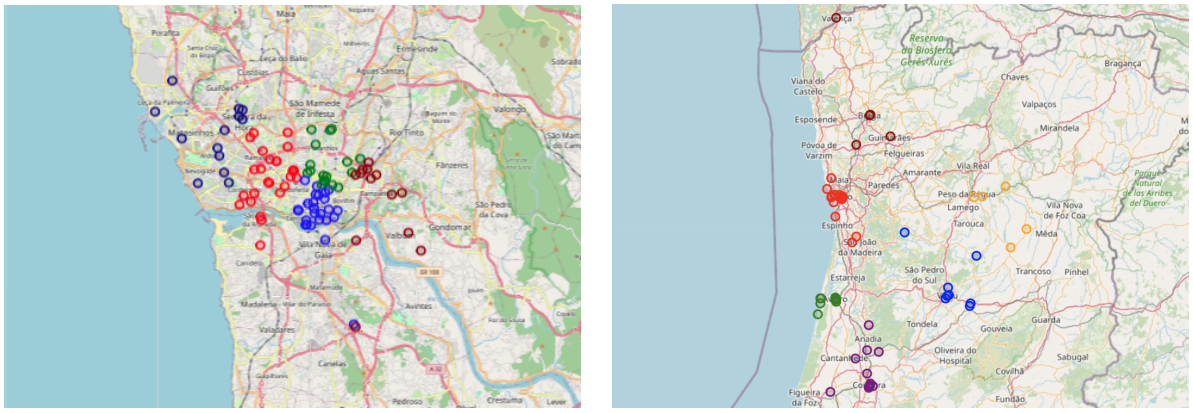

Figure 12: The two maps show the clusters that were formed from the taxi mobility data on the left image, and the POI data on the right image

Using K-Means, the set of location clusters C were derived based on the destinations and POIs. The result would be to assign each point to the closest centroid C_i. The clusters for the POI and taxi mobility coordinates were calculated first. After this step was completed, the Haversine distance between the cluster coordinates of the POI and each record on the taxi mobility data was computed. The final result was a cluster assigned to the dataset that provided a more accurate representation of the assigned cluster as it considered the POI as shown in image 13. In addition, some clusters ended up being consolidated, and the result produced three clusters that assigned the destination coordinates to three clusters.

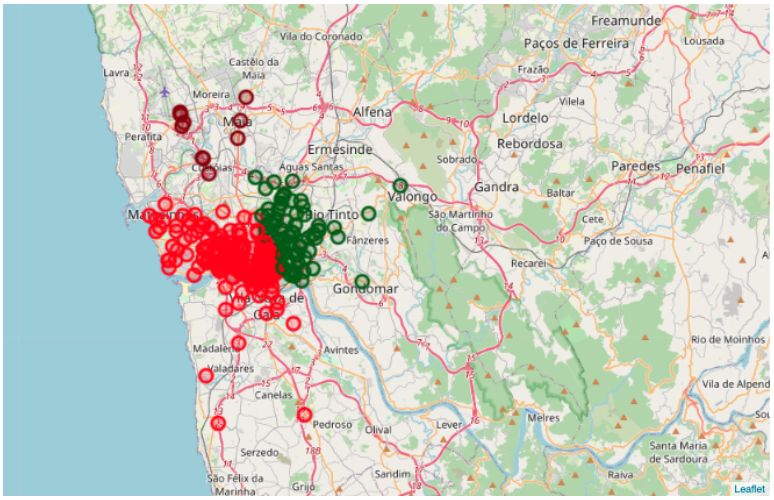

Figure 13: Destination points were assigned to the clusters that were derived from the POI data

4.5 Model Evaluation

In the model evaluation section, a comparison of all the different models is provided for both cases - with and without the POI data.

4.5.1 Basic Approach Without Points of Interest Data

A comparison of all the models was done for the basic approach as shown in table 4 and 3 to evaluate the model performance on both the training and test data. The main metrics, R-Square and Mean Square Error (MSE) are used to evaluate the model. R-Square is a common measure to determine how well the model fits the dependent variables. However, it does not take into consideration the

overfitting problem. If the regression model has many independent variables, because the model is too complicated, the R-Square may fit very well to the training data but performs badly for testing data. The MSE is also considered to measure the goodness of fit and compare different models. In comparing the models, the MSE is used to compare the models which results in the Gradient Boosting Regression as the best performing one from being the lowest number as shown in table 4.

Model	MSE Train Data	R^2 Train Data	CPU time
Gradient Boosting	0.000363	0.535672	876 ms
Histogram Gradient Boosting	0.000252	0.674832	1.82 s
Random Forest	0.000089	0.8844314	334 ms
k-Nearest Neighbors	0.000379	0.510322	102 ms
Decision Tree	0.0	1.0	103 ms

Table 3: Comparison of model performance on training data

Model	MSE Test Data	R^2 Test Data
Gradient Boosting	0.000479	0.425854
Histogram Gradient Boosting	0.000489	0.410312
Random Forest	0.000525	0.369489
k-Nearest Neighbors	0.000615	0.258115
Decision Tree	0.000915	-0.081319

Table 4: Comparison of model performance on test data

4.5.2 Basic Approach With Points of Interest Data

In this subsection, further details are provided to evaluate the exact same models in the previous approach with an addition of the POIs data as shown in table 8 and 5 to evaluate the model performance on both the training and test data. In comparing the models, the MSE is used to compare the models which results in the Random Forest Regression as the best performing one, and the Gradient Boosting Regression comes in second. It can be seen that all the models except for the Decision Tree Regression on the training data perform better when comparing the MSE in both the training and test datasets. The R-squared value also increased for all the models except for again the Decision Tree Regression on the training data. This could possibly suggest that model with the POI data fits better or it could be a symptom of an overfit model.

Model	MSE Train Data	R^2 Train Data	CPU time
Gradient Boosting	0.000138	0.805111	941 ms
Histogram Gradient Boosting	0.000087	0.876498	1.79 s
Random Forest	0.000003	0.955919	342 ms
k-Nearest Neighbors	0.000163	0.775128	105 ms
Decision Tree	0.0	1.0	67.2 ms

Table 5: Comparison of model performance on training data

Model	MSE Test Data	R^2 Test Data
Gradient Boosting	0.000280	0.672258
Histogram Gradient Boosting	0.000235	0.721299
Random Forest	0.000262	0.693549
k-Nearest Neighbors	0.000338	0.600404
Decision Tree	0.000374	0.564114

Table 6: Comparison of model performance on test data

5 Results

The methodology was tested on the test dataset from the Kaggle challenge. Even though the competition is over, it is still possible to submit an entry to see the score. When uploading the submission file, a score is obtained as shown in image 14. The two scores that are shown in image 14 show the public and private leaderboard score. The public leaderboard score is calculated on a subset of the testing set, whereas the final result will use the remaining data in the testing set to generate the private leaderboard score.

To evaluate the prediction error, the Mean Haversine Distance was used, as it was the evaluation metric used in the official Kaggle competition. Recall the section on Data Processing where it was discussed that the new Haversine feature was added to the model. From the perspective of the metrics such as the R-Squared and MSE, the model seems to improve. However, the model actually worsens on the leadership board when making uploading the submission files of the updated models to Kaggle as shown in the tables of the following subsections.

The K-Nearest Neighbors model as shown in 14 show a private and public score of 3.67182 and 3.64170 is obtained. This model is the paper's best-performing one on the private leaderboard. This score would place this study's model at rank 297 out of 381 teams. A score on one K-Nearest Neighbors model recorded by [Rossi et al., 2020] showed a score of 3.215 which is not that far from the the score of this model.

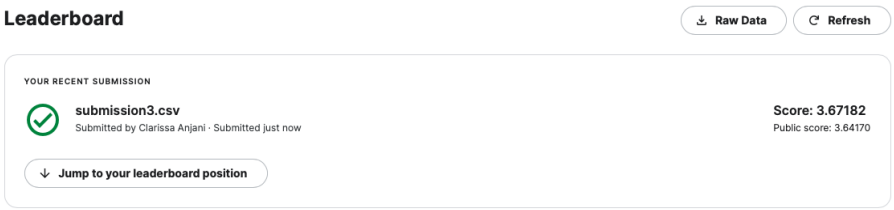

Figure 14: Score of best performing model in terms of Mean Haversine Distance on the private leaderboard is the K-Nearest Neighbors model

5.1 Basic Approach Without Points of Interest Data

This subsection examines the performance of all the regression models based on the approach without the POI data. On the public leaderboard, the Histogram-based Gradient Boosting model performs the best, and this would have placed the model at 301 out of 381 teams. Further details are shown in table 7.

Model	Private LB Score	R^2 Public LB Score
k-Nearest Neighbors	3.67182	3.64170
Histogram Gradient Boosting	3.71764	3.57366
Gradient Boosting	3.71764	3.57366
Random Forest	3.76196	3.67267
Decision Tree	3.71764	3.57366

Table 7: Comparison of model performance on test data ordered by the best performing private leaderboard score for the approach without the POI data

5.2 Basic Approach With Points of Interest Data

When reviewing the performance of all regression models, a comparison is made of the model on the data with the POI data. All models actually worsen except for the Histogram-based Gradient Boosting model that stays the same. The detailed comparison can be seen in table 8. .

Model	Private LB Score	R^2 Public LB Score
k-Nearest Neighbors	4.79756	4.46157
Histogram Gradient Boosting	3.71764	3.57366
Gradient Boosting	4.52343	4.29258
Random Forest	4.65065	4.43758
Decision Tree	4.52343	4.29258

Table 8: Comparison of model performance for the approach with the POI data

6 Alternative Approaches

In viewing the approaches on the Kaggle website, some approaches included non-machine learning approaches all the way to more complex models such as recurrent neural networks. The model of this paper was inspired by the existing literature available, while still considering the constraints of the class. Various simple to more complex approaches were tried by many others, that served as a basis for selecting the appropriate methodology of this paper for the purpose of understanding human movement patterns.

The challenge is normally represented as a multiclass classification problem, with the aim of selecting the next taxi destination form a set of known locations. In the realm of human mobility, the most straightforward way is to predict the next location over a number of POIs, and then treating the problem as a multiclass classification, that will then help in achieving the goal of predicting the GPS coordinates of the next cell [Rossi et al., 2020]. It was not uncommon to see people adopting clustering algorithms as well as is often the case when processing large data sets, see [Besse et al., 2018], [Rossi et al., 2020], [Ibrahim and Shafiq, 2018]. Another relatively straightforward approach is the K-Nearest neighbourhood. Given a pick-up point, it outputs the coordinates of the closest cluster centroid that considers the maximum clustering in a limited range, as well as the relationship between each of cluster center and its nearest neighbourhood, thus addressing the clustering issue of unevenly distributed datasets [Yu et al., 2021]. The approach should work quite well given that the Porto dataset is quite dispersed with many trips that occurred in far places in the city.

Another common way are using probabilistic methods in destination prediction. The probabilistic method matches the current trip with historical trajectories based on the ID of traveller, departure time, and departure location to determine the destination as the most fitting location that is matched [Tang et al., 2021]. However, to use this approach, deep learning methods need to be used to predict the destinations based on the trajectory points of the ongoing trip. The reason for this is to avoid the issue of data sparsity and such probabilistic matching methods will not work when the ongoing trip cannot be matched to historical trajectories which will likely happen. Such probabilistic approaches, treats the trajectory as a time series, and forecasts the destination based on the autocorrelation of the trajectory. In the Kaggle competition, the approach of [Besse et al., 2018] used a probabilistic method where the authors modelled the traffic flow patterns with a mixture of 2-D Gaussian distributions that resulted in a density-based clustering of locations. After this, the author assigns the new trajectory to the most appropriate cluster and then used the characteristics from the trajectories to predict the final destination.

The winning approach of the ECML/PKDD challenge used a Multi-Layer Perceptron (MLP) network trained on the taxi trajectory. It consisted of a hidden layer with 500 ReLU neurons and an output layer that is initialized with cluster coordinates. The network is trained with a stochastic gradient descent algorithm, a batch size of 200, a learning rate of 0.01, and a momentum term of 0.9. The model like other deep learning approaches, lack explainability. This deep learning approach allowed the usage of a lot of features to improve the model performance, whereas increasing the number of features in this paper's model actually worsened the performance.

Another promising approach that is similar to the approach of this study was done by [Rossi et al., 2020]. In [Rossi et al., 2020], a recurrent neural network is used which models the taxi drivers' behavior and encodes the semantics of the locations with a geographical information from the LSBNs. The authors also use a clustering algorithm like K-Means that assigns each POI to the closest cluster, such that for each point, a given mapped trajectory of clusters CT_i to leverage the related POIs.

A heavy clustering approach was done by [Ibrahim and Shafiq, 2018] where the authors clustered the trips using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN).

As previously mentioned that the authors in [Ibrahim and Shafiq, 2018] considered a K-Means clustering algorithm initially and moved to a DBSCAN algorithm. The authors claimed that it is more efficient in identifying the arbitrary shapes in the city's neighborhoods and dealing with noise trips. The authors also adopted a hierarchical clustering method called HDBSCAN that extended DBSCAN by constructing a tree of significant clusters.

Clustering in hotspot detection is significant for many applications such as city infrastructure construction, urban transportation planning and management, location-based service, and so on. In recent years, spatial clustering methods are widely used to discover hotspots from trajectory data. For instance, [Zhou et al., 2018] used K-Means clustering to analyze pick-up patterns of taxi service, and conducted location recommendation for taxis. Since K-Means seemed to have be a common approach done in literature in the realm of hotspot detection, it resulted in becoming a method that was worth attempting for the purpose of this paper.

7 Conclusion

This paper has introduced another way of looking at the topic of next destination prediction of taxis by integrating human mobility patterns. Instead of using only the trajectory data, the methodology relies on data of POIs. With POIs, it is expected to provide a more wholesome picture of transportation routes with some aspect of human motivations.

In the methodology, version 3 of the Foursquare API was used, which was never done before as the competition took place five years ago where the participants were still using version 2 of the API. There were limited resources available on how to query and structure the data of version 3 of the Foursquare API, but I was able to extract the needed data which can be seen on the Google Colab file on the Foursquare data.

The results show that the models exhibit comparatively low performance for both prediction tasks, with more time, some further improvements could be made. It should however be noted that the attempted K-Nearest Neighbors model in this paper produced a private and public leaderboard score of 3.67182 and 3.64170 which is not so far off from the recorded K-Nearest Neighbors score of 3.215 that was recorded by [Rossi et al., 2020]. While the results may not show concrete improvement when integrating the POI data, it is still worthwhile to examine the field of integrating taxi mobility data with POI data.

Nevertheless on a personal level, it was beneficial to study the taxi trips in Porto as it was possible to discover more attractions in this beautiful city and it was interesting to see analyze people's. movements and top places of interest that were high on the list of places to visit.

7.1 Limitations

From the model results, the results on the leaderboard worsen when it based it the model is measured against the official metric of the challenge, the Mean Haversine distance. Various factors could contribute to this.

Firstly, the POI data that was used only provided details of 67 POIs when others were able to query about 8K POIs with Version 2 of the Foursquare Places API. Version 2 of the Foursquare API was deprecated and has been widely used in research by others like [Rossi et al., 2020] and [Ibrahim and Shafiq, 2018]. New implementations can no longer use this version of the API and need to either use version 3 of the API or perhaps the flat file. This study used version 3 of the Foursquare Places API which only generated a minimal number of POI, as compared to the thousands that were generated in other approaches, so it would be valuable to research alternative ways of extracting a significantly larger number of POIs, either by using another approach from Foursquare (i.e. the flat file), or another API provider for providing such LSBN data such as data on Twitter activity or phone usage. In this approach, the Foursquare Places API version 3 was used because new implementations cannot use the version 2 of the Places API. Foursquare also provides a flat file option that could have provided more data on the POIs, but unfortunately time constraints didn't allow further research into implementing the flat file. The flat file could potentially provide more than 67 POIs.

The current score of the model is much lower than the other participants of the leaderboard, perhaps due to the comparatively simple approach. It would also be interesting to try more advanced approaches to improve the score by doing more complicated deep learning or feature engineering to

improve the performance. In addition, this paper implemented a simple clustering algorithm with 6 to 7 clusters, whereas others used clusters from forty five to several thousand clusters. Perhaps it might be worthwhile to also use a hierarchical clustering approach to account for the categories of the places as was done in [Rossi et al., 2020]. The authors used hierarchical clustering to take into account the categories (e.g. Professional and Other Places → Medical Center → Doctor's office to produce a more informative representation of the area. Since this study uses a rather simple clustering approach compared to other complex clustering methods as mentioned in the use of HDBSCAN clustering in [Ibrahim and Shafiq, 2018] or the Mean-Shift Clustering approach in [Tang et al., 2021], it could be worthwhile to switch to other clustering algorithms. Lastly, further analysis needs to be done on whether the model has issues with overfitting especially for cases when the R-squared is rather high, above 0.7. Various measures could be added such as including the adjusted R-Squared.

Other Kaggle challenge participants not only examined the aspect of next destination prediction, they also included trip duration in the features. While the timestamp was included in the exploratory data analysis, more could have been done with this data.

7.2 Further proposed steps

Considering the limitations and interesting areas, it would be worthwhile to extend the approach with the some additional insights. Extending this analysis can predict future routes and places which can build a useful recommendation system for drivers and challenges. An interesting problem is he optimization of the total profit guaranteed by the entire company fleet that consists of the definition of a comprehensive mobility pattern to support drivers to decide which of two potential rides will be more profitable [Rossi et al., 2020]. There is still much room for improvement in the area of driver recommendation systems, and having such systems can optimize the operations of taxis and drivers profits. In addition, because taxi driver fraud is a significant problem, it could also be useful to extend the approach to developing a taxi monitoring system and evaluate if unethical taxi drivers detoured customers away from their original location [Zhou et al., 2018].

As taxis contribute to urban traffic, the topic of traffic congestion and unkept roads could be an interesting extension. If traffic congestion can be identified from observing which taxis have been on the road at a specific time for too long, it could be interesting to see if traffic congestion is really due to unkept roads to direct urban planners to perform road maintenance and repairs in that location [Rossi et al., 2020]. As different cities have different road network structures, traffic policies and management measures, the model didn't consider such additional complexities. In addition, different modes of transportation also have different traffic patterns, so combining the different data sources could be considered to understand the impact of different transportation modes, road network structures, and transportation policies on the taxi destination prediction. This could then result in an even better traffic management and control.

References

[Kag, a] Ecml/pkdd 15: Taxi trajectory prediction (i). https://www.kaggle.com/competitions/pkdd-15-predict-taxi-service-trajectory-i/overview. Accessed: 2022-06-28.

[Kag, b] How to use the kaggle api from colab. https://colab.research.google.com/github/corrieann/kaggle/blob/master/kaggle_api_in_colab.ipynb#scrollTo=hMY4CFezjcG-. Accessed: 2022-06-28.

[Kag, c] Nyc taxi trip prediction using decision tree. https://www.kaggle.com/code/abhi8901/nyc-taxi-trip-prediction-using-decision-tree/notebook. Accessed: 2022-06-28.

[Rpu,] Predicting taxi farew with random forests. https://rpubs.com/eichemberger/791065. Accessed: 2022-06-28.

[Skl, a] sklearn.ensemble.decisiontreeregressor. https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html. Accessed: 2022-06-28.

[Skl, b] sklearn.ensemble.gradientboostingregressor. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html. Accessed: 2022-06-28.

[Skl, c] sklearn.ensemble.histgradientboostingregressor. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.HistGradientBoostingRegressor.html. Accessed: 2022-06-28.

[Skl, d] sklearn.neighbors.kneighborsregressor. https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsRegressor.html. Accessed: 2022-06-28.

[Kag, d] Taxi trajectory prediction i. https://www.kaggle.com/code/dahouda/taxi-trajectory-prediction-i. Accessed: 2022-06-28.

[Por, 2020] (2020). Portugal is best european destination for the 4th year running and porto is world's leading city break 2020. https://www.porto.pt/en/news/portugal-is-best-european-destination-and-porto-is-worlds-leading-city-break-destination-2020-for- Accessed: 2022-06-28.

[Ahmed et al., 2008] Ahmed, Q. I., Lu, H., and Ye, S. (2008). Urban transportation and equity: A case study of beijing and karachi. *Transportation Research Part A: Policy and Practice*, 42(1):125–139.

[Besse et al., 2018] Besse, P. C., Guillouet, B., Loubes, J.-M., and Royer, F. (2018). Destination prediction by trajectory distribution-based model. *IEEE Transactions on Intelligent Transportation Systems*, 19(8):2470–2481.

[Chen et al., 2009] Chen, C., Chen, J., and Barry, J. (2009). Diurnal pattern of transit ridership: a case study of the new york city subway system. *Journal of Transport Geography - J TRANSP GEOGR*, 17:176–186.

[Chen et al., 2014] Chen, C., Zhang, D., Li, N., and Zhou, Z.-H. (2014). B-planner: Planning bidirectional night bus routes using large-scale taxi gps traces. *IEEE Transactions on Intelligent Transportation Systems*, 15.

[da Silva and Fileto, 2015] da Silva, F. P. and Fileto, R. (2015). A method to detect and classify inconsistencies of moving objects' stops with requested and reported tasks. *Journal of Information and Data Management*, 6(1):71.

[Gade, 2010] Gade, K. (2010). A non-singular horizontal position representation. *Journal of Navigation*, 63(3):395–417.

[Gallotti et al., 2015] Gallotti, R., Bazzani, A., Rambaldi, S., and Barthelemy, M. (2015). How transportation hierarchy shapes human mobility.

[Hägerstrand, 1970] Hägerstrand, T. (1970). What about people in regional science? *Papers of the Regional Science Association*, 24:6–21.

[Ibrahim and Shafiq, 2018] Ibrahim, R. and Shafiq, M. O. (2018). Mining trajectory data and identifying patterns for taxi movement trips. In *2018 Thirteenth International Conference on Digital Information Management (ICDIM)*, pages 130–135.

[Kang and Qin, 2016] Kang, C. and Qin, K. (2016). Understanding operation behaviors of taxicabs in cities by matrix factorization. *Computers Environment and Urban Systems*, 60:79–88.

[Lam et al., 2015] Lam, H. T., Diaz-Aviles, E., Pascale, A., Gkoufas, Y., and Chen, B. (2015). (blue) taxi destination and trip time prediction from partial trajectories.

[Li et al., 2011] Li, B., Zhang, D., Sun, L., Chen, C., Li, S., Qi, G., and Yang, Q. (2011). Hunting or waiting? discovering passenger-finding strategies from a large-scale real-world taxi dataset. In *2011 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, pages 63–68.

[Liu et al., 2012] Liu, Y., Kang, C., Gao, S., Xiao, Y., and Tian, Y. (2012). Understanding intra-urban trip patterns from taxi trajectory data. *Journal of Geographical Systems*, 14(4):463–483.

[Liu et al., 2012] Liu, Y., Wang, F., Xiao, Y., and Gao, S. (2012). Urban land uses and traffic 'source-sink areas': Evidence from gps-enabled taxi data in shanghai. *Landscape and Urban Planning*, 106(1):73–87.

[Peng et al., 2012] Peng, C., Jin, X., Wong, K.-C., Shi, M., and Liò, P. (2012). Correction: Collective human mobility pattern from taxi trips in urban area. *PLOS ONE*, 7(8):null.

[Phiboonbanakit and Horanont, 2016] Phiboonbanakit, T. and Horanont, T. (2016). How does taxi driver behavior impact their profit? discerning the real driving from large scale gps traces. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, UbiComp '16, page 1390–1398, New York, NY, USA. Association for Computing Machinery.

[Rossi et al., 2020] Rossi, A., Barlacchi, G., Bianchini, M., and Lepri, B. (2020). Modelling taxi drivers' behaviour for the next destination prediction. *IEEE Transactions on Intelligent Transportation Systems*, 21(7):2980–2989.

[Smith et al., 2017] Smith, A., Kun, A., and Krumm, J. (2017). Predicting taxi pickups in cities: which data sources should we use? pages 380–387.

[Tang et al., 2021] Tang, J., Liang, J., Yu, T., Xiong, Y., and Zeng, G. (2021). Trip destination prediction based on a deep integration network by fusing multiple features from taxi trajectories. *IET Intelligent Transport Systems*, 15.

[Xu et al., 2017] Xu, M., Zhang, L., LI, W., and Xia, H. (2017). Mobility pattern of taxi passengers at intra-urban scale:empirical study of three cities. *()*, 5:537–555.

[Yu et al., 2021] Yu, Q., Chen, C., Sun, L., and Zheng, X. (2021). Urban hotspot area detection using nearest-neighborhood-related quality clustering on taxi trajectory data. *ISPRS International Journal of Geo-Information*, 10:473.

[Yuan et al., 2011] Yuan, J., Zheng, Y., Zhang, L., Xie, X., and Sun, G. (2011). Where to find my next passenger. UbiComp '11, page 109–118, New York, NY, USA. Association for Computing Machinery.

[Yuan et al., 2012] Yuan, N., Zheng, Y., Zhang, L., and Xie, X. (2012). T-finder: A recommender system for finding passengers and vacant taxis. *Knowledge and Data Engineering, IEEE Transactions on*, 25.

[Zhang et al., 2019] Zhang, H., Shi, B., Zhuge, C., and Wang, W. (2019). Detecting Taxi Travel Patterns using GPS Trajectory Data, A Case Study of Beijing. *KSCE Journal of Civil Engineering*, 23(4):1797–1805.

[Zhao et al., 2015] Zhao, P., Qin, K., Zhou, Q., Liu, C., and Chen, Y. (2015). Detecting hotspots from taxi trajectory data using spatial cluster analysis. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, II-4/W2:131–135.

[Zheng et al., 2013] Zheng, Y., Liu, F., and Hsieh, H.-P. (2013). U-air: When urban air quality inference meets big data. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, page 1436–1444. Association for Computing Machinery.

[Zhou et al., 2018] Zhou, D., Hong, R., and Xia, J. (2018). Identification of taxi pick-up and drop-off hotspots using the density-based spatial clustering method. pages 196–204.

[Zhou et al., 2016] Zhou, Z., Dou, W., Jia, G., Hu, C., Xu, X., Wu, X., and Pan, J. (2016). A method for real-time trajectory monitoring to improve taxi service using gps big data. *Information and Management*, 53(8):964–977. Big Data Commerce.