Sensing Urban Mobility with Taxi Flow

Conference Paper · November 2011

DOI: 10.1145/2063212.2063215

3 authors:

Marco Veloso
Instituto Politécnico de Coimbra
36 PUBLICATIONS **182** CITATIONS
SEE PROFILE

Santi Phithakkitnukoon
Chiang Mai University
80 PUBLICATIONS **695** CITATIONS
SEE PROFILE

Carlos Lisboa Bento
University of Coimbra
121 PUBLICATIONS **648** CITATIONS
SEE PROFILE

Sensing Urban Mobility with Taxi Flow

Marco Veloso
Centro de Informática e Sistemas da Universidade de Coimbra, Portugal
Escola Superior de Tecnologia e Gestão de Oliveira do Hospital, Portugal
mveloso@dei.uc.pt

Santi Phithakkitnukoon
Culture Lab, School of Computing Science, Newcastle University, United Kingdom
SENSE*able* City Lab, Massachusetts Institute of Technology, Cambridge, MA, USA
santi@mit.edu

Carlos Bento
Centro de Informática e Sistemas da Universidade de Coimbra, Portugal
bento@dei.uc.pt

ABSTRACT
The analysis of taxi flow can help better understand the urban mobility. In this work, we analyze 177,169 taxi trips collected in Lisbon, Portugal, to explore the relationships between pick-up and drop-off locations; the behavior between the previous drop-off to the following pick-up; and the impact of area type in taxi services. We also carry out the analysis of predictability of taxi trips given history of taxi flow in time and space.

Categories and Subject Descriptors
I.5.2. Patter Recognition: Pattern analysis

General Terms
Algorithms.

Keywords
Urban mobility, spatiotemporal analysis, taxi-GPS traces, naïve Bayesian classifier.

1. INTRODUCTION
With the development of pervasive technologies (e.g. Global Positioning System, Global System for Mobile Communications), the urban areas are meeting the vision of smart cities, where sensors combine with the environment to perceive the current status and interact accordingly.

However, the urban areas are experiencing a growth in size and population. The constant movement of people demands for changes in the transportation systems, to improve public transportation modes (e.g. bus, metro, train) in order to meet citizens needs and therefore reduce the use of individual means of transport (e.g. car). Thus, an efficient public transportation system can lead to more efficient environmental urban areas, since there is a reduction of traffic congestions and consequent energy consumption.

To optimize the public transportation network it is essential to understand what drives the common citizen, what their needs are. Retrieving data from the traditional public transportation (e.g. bus, train, metro) can provide a relevant database of samples and general passengers' movement. However, does not provide the exact origin and destination for each passenger, since these transportation modes rely on pre-designated stops and paths. The taxi service can be a way to retrieve large dataset of information with a higher precision when we focus the origin and destination of each trip. It can pick-up the passengers right where they are standing, and then drop-off precisely in the desirable destination, without being bounded to a pre-determined path. The process of data collecting is transparent and non-intrusive to the passenger.

Our on-going work is focused on the analysis of taxi-GPS traces acquired in the city of Lisbon, Portugal, to better understand urban mobility. The contribution of this work lies on the following two aspects: spatiotemporal analysis and study of predictability of taxi trips. For the former, we analyze taxi traces to identify the relationships between pick-up and drop-off locations; characterize the scenario between taxi services (i.e. what happens between the latest drop-off and next pick-up) in order to improve taxi profit; and explore the value of Points of Interest in analysis of taxi flow. For the latter, we explore the possibility of predicting the next destination given hour of the day, day of the week, weather condition, and area type.

2. RELATED WORK
Liu et al. [1] classify taxi drivers into the top and standard drivers according to their income. Based on 3,000 taxi drivers, they observe that top drivers have the special proportion of operation zones, with an optimal balance between taxi travel demand and fluid traffic conditions, while ordinary drivers operate in fixed spots with few variations.

Ziebart et al. [2] present a decision modeling framework for probabilistic reasoning from observed context-sensitive actions. Based on 25 taxi drivers, the model is able to make decisions regarding intersections, route, and destination prediction given partially traveled routes.

Yuan et al. [3] and Zheng et al. [4] propose the T-Drive system that relies on an historical GPS dataset generated by over 33,000 taxis in a period of three months, to present the algorithm to compute the fastest path for a given destination and departure time. Zheng et al. also describe a three-layer architecture with the notion of landmark graph to model the knowledge of taxi drivers.

Chang et al. [5] propose a four-step approach for mining historical data in order to predict demand distributions considering time, weather, and taxi location. They show that different clustering methods have different performances on distinct data distributions.

Phithakkitnukoon et al. [6] present a model to predict the number of vacant taxis for a given area of the city using a naïve Bayesian classier with developed error-based learning algorithm and mechanism for detecting adequacy of historical data. With 150 taxi drivers, they achieve an overall error rate of less than one taxi per 1x1 km^2 area.

Qi et al. [8] investigates the relationship between regional pick-up and drop-off characteristics of taxi passengers and the social function of city regions. They develop a simple classification method to recognize regions' social function which can be break in Scenic Spots, Entertainment Districts and Train/Coach Stations.

There are also studies performed by Yang et al. [9] and Wong et al. [10] in order to improve the taxi service in congestion scenarios.

3. SPATIOTEMPORAL ANALYSIS

The exploratory analysis is a useful tool to identify emerging patterns and obtain a better understanding of the variables that model the system. In this section, we will describe the source dataset, previous work and explore the following aspects: spatial relationships between pick-up and drop-off locations; analysis of the movement of taxis between services (i.e. from the previous drop-off to the following pick-up); and the impact of area type characterized by Points of Interest (POIs) to the taxi service.

3.1 Dataset

For the present study we use a database with more than 10 million taxi-GPS samples from August through December in 2009, collected in Lisbon, Portugal by GeoTaxi [11]. For study purposes, only pick-up and drop-off locations and timestamps are considered, which correspond to 177,169 distinct trips. A data cleaning process was applied, removing trips with less than 200m and more than 30km. Data was collected from 217 distinct taxis, which account for nearly 15% of taxis in Lisbon area.

Weather conditions for the period under study were retrieved from Weather Underground [12] and a collection of 10,954 Points Of Interest (POI), grouped into eight categories (Services 16.96%, Recreation 14.78%, Education 20.84%, Shopping 4.65%, Police 2.81%, Health facilities 6.58%, Transportation 2.28%, Accommodation 1.81%), was provided by Sapo Maps [13].

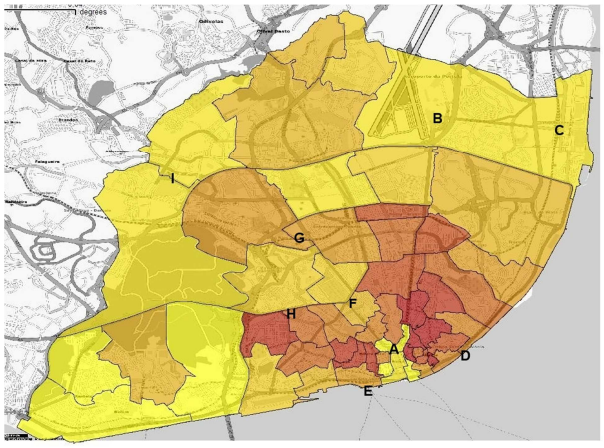


Figure 1. Lisbon council and population density (A, City downtown; B, Airport; C, Train Station; D, Train Station; E, Ferry dock; F, City center; G, Univ. Campus; H, Commercial Area; I, Residential).

LBSN'10, November 1, 2010, Chicago, Illinois, USA.

The area of study encompasses the Lisbon council (figure 1) that consists of 53 parishes, an area of around 110 km^2, and a population of 800,000 habitants. The city downtown is the central area, which includes the oldest and smallest parishes with greatest population density, touristic, historic and commercial areas, and the interface for several public transportation services (bus, metro, train and ferry). Moving from the city center there are larger area parishes with lower population density, which are characterized by residential areas surrounding business areas. Major infrastructures (e.g. airport, industrial facilities) are located in the city's periphery.

For the analysis, we model the Lisbon map with grids of 0.5x0.5 km^2.

3.2 Previous work

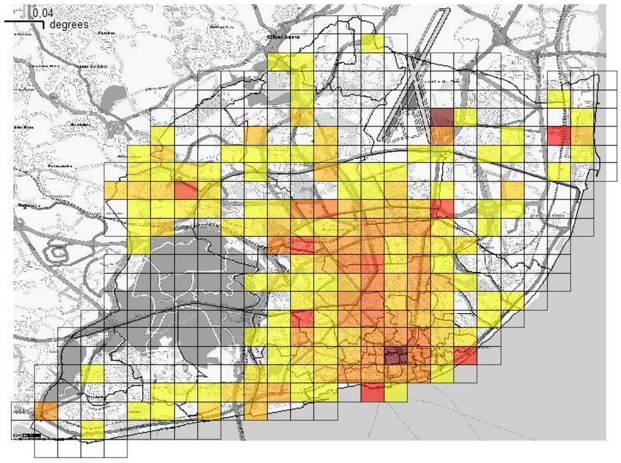
In previous works ([6], [7]), the same data was applied to better understand and predict the use of taxi service. The former focused on predicting the number of vacant taxis in the city. A predictive model was proposed based on the naïve Bayesian classifier using the error-based learning algorithm. The average error of the system ranges from one (in locations with lower density of taxi services) to three vacant taxis (locations with higher density of taxi services). Using the information theory's mutual information has been shown to detect the adequacy of historical data and hence reduce computational cost of the proposed model.

The latter presented an exploratory analysis of the spatiotemporal distribution of taxi pick-ups and drop-offs, identifying the most frequent locations (e.g. city downtown, airport, and major train stations) and temporal periods (an increase from 7 a.m., reaching the pick at 12 p.m., and a slow decreasing from 2 p.m.). Exponential distribution was observed for the trip duration and revenue while Gama distribution was observed for the trip distance. The strategies of the taxi drivers were also presented. The majority of the taxi drivers used combined strategies, mainly driving around the city and in certain time periods staying at a fixed location (e.g. at the airport from 6 a.m. to 8 a.m.).

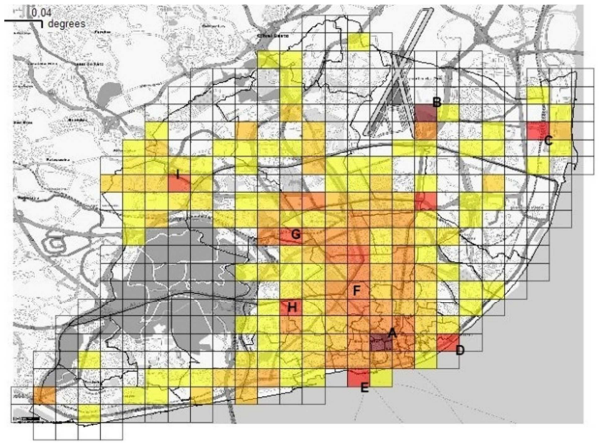
In contrast to the previous work, here we are presenting an exploratory analysis focused in the relationship between pick-up and drop-off locations; the movement of taxis between services; and the impact of spatial profile characterized by POIs to the taxi service. We also present the analysis of predictability of taxi trips.

3.3 Spatial relationships between pick-up and drop-off locations

The overall taxi service distribution in Lisbon is depicted in figure 2 where some major locations are identified, such as city downtown (A), airport (B), train stations (C, D) and ferry dock (E).



Pick-up locations.



Drop-off locations.

Figure 2. Taxi pick-up (top) and drop-off locations (bottom) density (A, City downtown; B, Airport; C, Train Station; D, Train Station; E, Ferry dock; F, City center; G, Univ. Campus; H, Commercial Area; I, Residential).

In order to understand how the pick-up and drop-off location areas relate we compute the number of trips between every two possible locations. The result is shown in figure 3 where the thickness of the line represents this intensity. Strong relations can be observed in links B-C, D-E, D-A, A-F, and F-B. All those locations are characterized by some public transportation modality. B is the access to the airport, C and D are trains stations, E is a ferry dock, A and F are bus stops zones. From this observation, we hypothesize that the taxi service is often used as a bridge between public transportation modalities. It is also important to point out that the locations A, C and F (some of the most frequent pick-up or drop-off locations) give access to services and commercial areas.

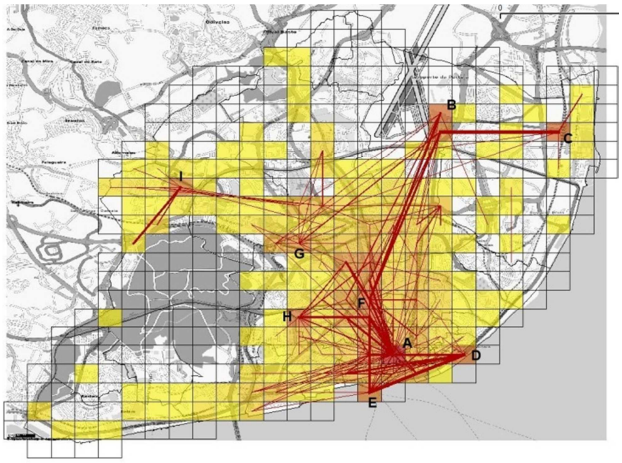


Figure 3. How strongly connected locations are, according to taxi services (A, City downtown; B, Airport; C, Train Station; D, Train Station; E, Ferry dock; F, City center; G, Univ. Campus; H, Commercial Area; I, Residential).

3.4 Downtime analysis

Another possibility is that the taxi drivers may want to improve their income by targeting the above-mentioned locations. To better understand that we need to consider what happens in between services (i.e. downtime – time spent looking for next pick-up).

Figure 4 shows the variation in trips made by and the number of taxis in service throughout the day. Figure 5 shows the average time spent and distance traveled during downtime. In the early AM hours (12 a.m. to 7 a.m.), due to the low amount of taxis in service, the average downtime and distance traveled searching for new passengers are relatively high. The average downtime remains almost constant during 10 a.m. to 10 p.m. There is a sudden drop in downtime at 10 p.m. but a rise of distance traveled. The lower number of taxis in service as well as potential passengers during this late hour presumably causes longer time spent searching for pick-up. Both distance traveled and downtime appear to follow exponential distributions – as shown in figure 6.

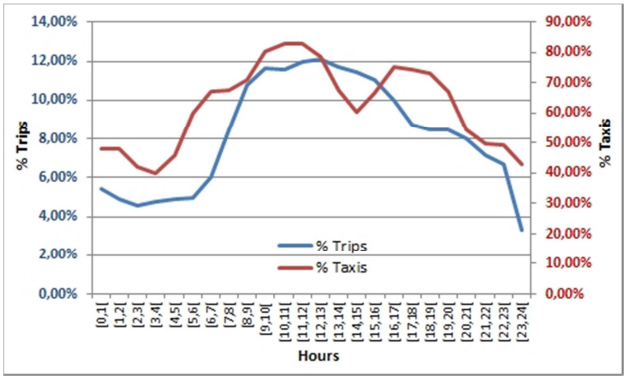


Figure 4 Amount of trips made by (blue) and number of taxis in service (red) throughout the day.

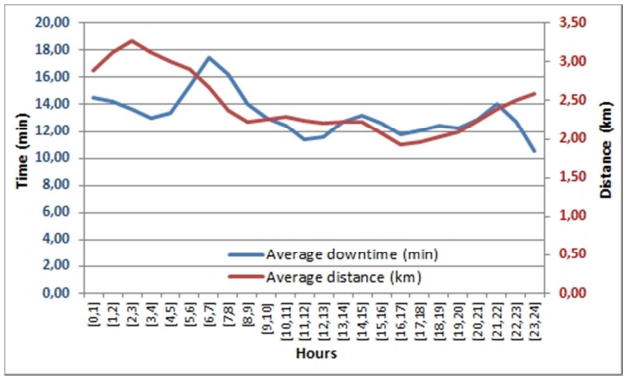


Figure 5. Average downtime (blue) and distance traveled (red).

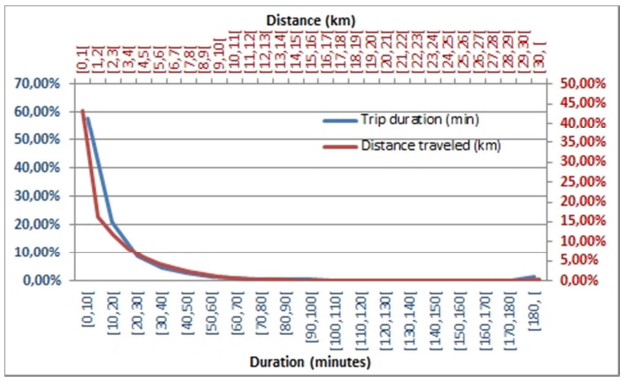


Figure 6. Distribution of distance traveled (red) and time spent (blue) during downtime.

In figure 7, we can see the relationship between the distance traveled during downtime, and the resulting service distance with corresponding average income. A higher distance traveled during downtime does not guarantee a more profitable service.

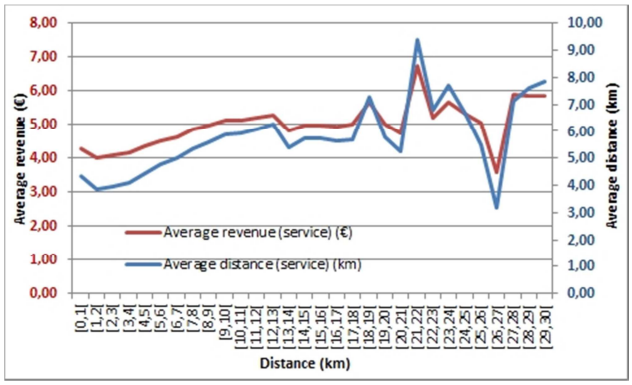


Figure 7. Variation in service distance (blue) and income (red).

Figure 8 presents the areas with high (red) and low (yellow) average distance traveled when taxis search for new pick-ups and the relationship between the previous drop-off locations and the following pick-up locations. The areas away from the city center (characterized by higher residential density) show higher average distances traveled between services, whereas in downtown the distances traveled are relatively smaller. By the same token, strong relationships between adjacent locations are observed in urban areas while in suburban areas strong links are observed between distant locations. This appears to us that after a drop-off in suburban area, a taxi driver typically heads to locations with higher probability of picking up new passengers (e.g. airport, city center) even if it means to travel a higher distance to the next pick-up location.

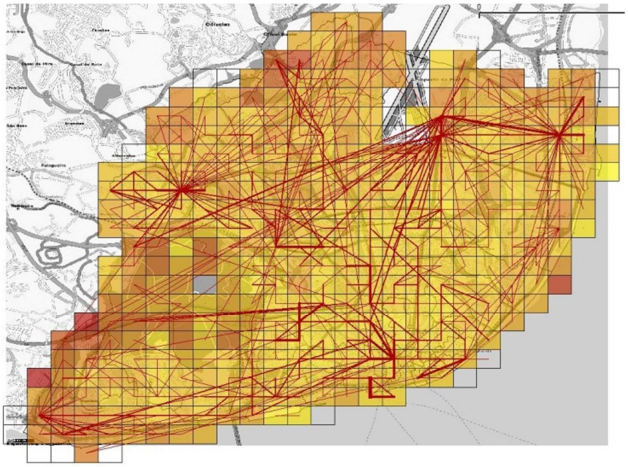


Figure 8. Spatial distribution according to the average distance traveled during downtime (red corresponds to high value) and the relationship between previous drop-off and next pick-up location (line thickness represents strength).

As an overall observation, in order to improve the profit, it is preferable for a taxi driver to wait for passengers in locations related with main public transportation terminals, and not travel great distances to the next pick-up location, unless to return to the aforementioned locations. If the drop-off location coincides with a public transportation terminal it is preferable to wait for new passengers in that location.

3.5 Impact of area type characterized by POIs to the taxi service

By embedding spatial profile like Points of Interest (POIs) onto the map, we can further observe taxi dynamics according to the area characteristic.

Figure 9 shows the map with POIs that are grouped in eight different categories, and figure 10 shows the distribution of these categories. One can observe that Education facilities (e.g. kindergarten, university, etc.), Recreation (bar, restaurant, etc.) and Services (e.g. bank, etc.) are the dominant POI categories (which account for over 70%).

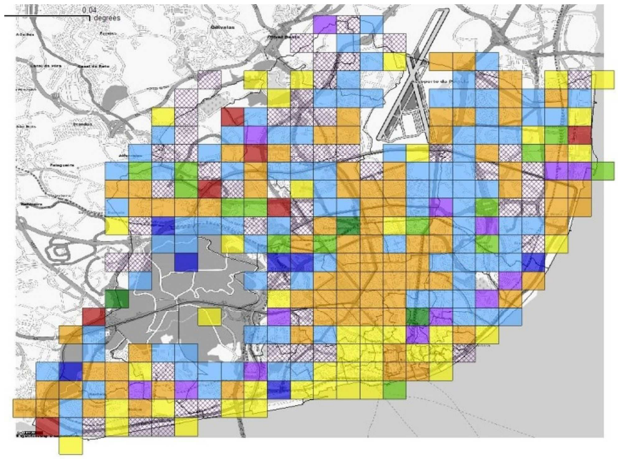


Figure 9. Predominant POI category on each location (colors correspond to classification performed in figure 9).

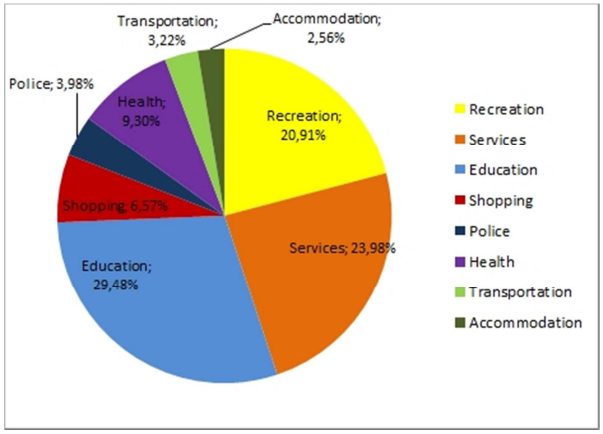


Figure 10. POIs categories distribution.

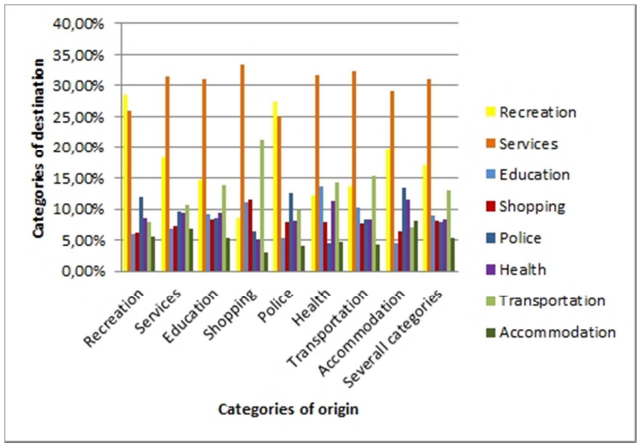


Figure 11. Origin-destination distribution according to area type characterized by POI categories.

This POI map allows us to further explore the origin-destination area type. Figure 11 shows that Services and Recreation are the most frequent drop-off area types – independent of the pick-up location. Transportation is the most likely drop-off area if the pick-up is Shopping, Transportation, Health, or Education. Education, as the most likely drop-off area, is mainly connected to pick-up locations from Health area. This observation is an indication that the area type can be a possible predictive attribute to consider when looking for taxi demand.

4. PREDICTABILITY ANALYSIS

Contrary to other public transportation modes, the taxi movement dynamically adapts to the flow and the need of the city. In previous work [6] we carried out a spatiotemporal analysis of trips made by taxis and found that day of the week, time of the day, and weather condition are promising features in predicting taxi volume. In this work, we aim to explore the predictability of taxis given the current drop-off location. We have observed that area type characterize by POI can potentially be used here along with other aforementioned features used in the previous work. Here we apply a simple probabilistic approach.

We apply a naïve Bayesian classifier for our study of the predictability. The classifier simply applies the Bayes' theorem with independence assumption [14]. The objective is to compute the likelihood of each possible grid cell destination (Y) given the hour of the day (T), day of the week (D), weather condition (W) and area type (I). The conditional probability can be formulated as follows:

$$P(Y = y_i|T, D, W, I) = \frac{P(Y = y_i)P(T, D, W, I|Y = y_i)}{P(T, D, W, I)} \quad (1)$$

where T = {1, 2, …, 24}, D = {Sunday, …, Saturday}, W = {Sunny, Cloudy, Rainy}, and I = {Services, Recreation, Education, Shopping, Police, Health, Transportation, Accommodation}. The prediction is based on the *maximum a posteriori probability* (MAP) decision rule:

$$\begin{aligned} y_{MAP} &= \arg\max_{y_i \in Y} P(Y = y_i|T, D, W, I) \\ &= \arg\max_{y_i \in Y} P(Y = y_i)P(T, D, W, I|Y = y_i) \\ &= \arg\max_{y_i \in Y} P(Y = y_i) \\ &\quad \prod_i P(T|Y = y_i)P(D|Y = y_i)\,P(W|Y = y_i)P(I|Y = y_i) \end{aligned} \quad (2)$$

Based on 5-folds cross validation, we are able to predict (for each pick-up) the next destination at about 5%. To further investigate on this predictability aspect, we examine the *predicted list* [15] – the list of the most likely destinations where the top of the list contains more likely destinations than the ones lower on the list.

Figure 12 shows that accuracy rate varying with the length of the predicted list. The list must grow to about 90 possible destinations in order to predict the correct destination at a high accuracy (70%). This reflects the randomness of the taxi flow in the city. Although the previous work [6] has shown a promising result in predicting vacant taxi volume at a large scale, predicting taxi individually shown in this work here appears to be rather difficult.

If we modify the objective to compute the likelihood of each possible area type destination given the same variable (hour of the day, day of the week, weather conditions and area type) we are

able to predict the next area type destination at about 47%, an expected improvement since we reduced the class domain.

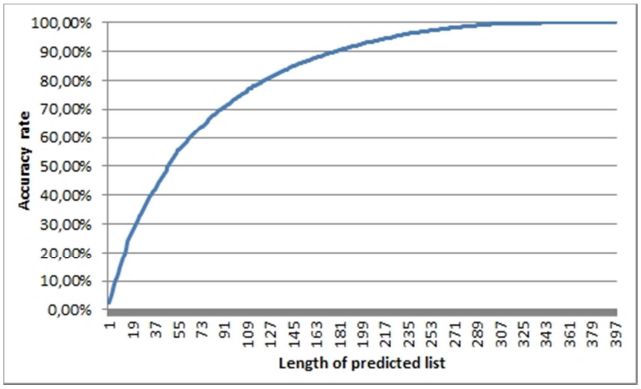


Figure 12. Corresponding accuracy rate for growing length of the predicted list.

5. CONCLUSIONS

By analyzing the taxi-GPS traces from Lisbon, Portugal, we are able to identify the link between pick-up and drop-off locations, the behavior during downtime – time spent searching for next pick-ups – and the impact of area type characterized by POIs to the taxi service. We observed strong links between public transportation terminals and taxis tend to avoid making long trips to suburban areas for pick-up. We also verified that Transportation is the most likely drop-off area if the pick-up is Shopping, Transportation, Health, or Education, and Education is the most likely drop-off area if the pick-up Health area.

Our predictability analysis shows that individual taxi trips are relatively random. With Bayesian approach given time of the day, day of the week, weather condition, area type, and the current pick-up location, only 5% of all trips are predictable. Being able to accurately predict taxi flow is important and a challenging problem, which we will address it further in our future work. Other topics for our future studies include the commuting pattern between multimodality as suggest by the exploratory analysis.

6. REFERENCES

[1] Liu, L., Andris, C., Bidderman, A., Ratti, C.: Revealing taxi drivers mobility intelligence through his trace. Movement-Aware Applications for Sustainable Mobility: Technologies and Approaches, (2010), 105-120.

[2] Ziebart, B.D., Maas, A.L., Dey, A.K., Bagnell, J.A.: Navigate like a cabbie: probabilistic reasoning from observed context-aware behavior. In: UbiComp '08: Proc. of the 10th int. conf. on Ubiquitous computing, New York, NY, USA, ACM (2008), 322-331.

[3] Yuan, J., Zheng, Y., Zhang, C., Xie, W., Xie, X., Huang, Y.:T-Drive: Driving Directions Based on Taxi Trajectories, in Proc. ACM SIGSPATIAL GIS 2010, Association for Computing Machinery, Inc. 1 (2010), 99-108.

[4] Zheng, Y., Yuan, J., Xie, W., Xie, X., Sun, G.: Drive Smartly as a Taxi Driver. In 7th Int. Conference on Ubiquitous Intelligence & Computing and 7th Int. Conference on Autonomic & Trusted Computing (UIC/ATC) (2010), 484-486.

[5] Chang, H., Tai, Y., Hsu, J.Y.: Context-aware taxi demand hotspots prediction. Int. J. Bus. Intell. Data Min. 5(1) (2010), 3-18.

[6] Phithakkitnukoon, S., Veloso, M., Bento, C., Biderman, A., Ratti, C.: Taxi-Aware Map: Identifying and predicting vacant taxis in the city. In Proc. AmI 2010, First International Joint Conference on Ambient Intelligence (2010), 86-95.

[7] Veloso, M., Phithakkitnukoon, S., Bento, C., Olivier, P., Fonseca, N,: Exploratory Study of Urban Flow using Taxi Traces. In First Workshop on Pervasive Urban Applications (PURBA) in conjunction with Pervasive Computing, San Francisco, California, USA, (2011).

[8] Qi, G., Li, X., Li, S., Pan, G., Wang, Z., Zhang, D., Measuring Social Functions of City Regions from Large-scale Taxi Behaviors. In PerCom- Workshops 2011, pp. 21-25, Seattle, USA, (2011).

[9] Yang, H., Ye, M., Tang, W.H., Wong, S.C.: Regulating taxi services in the presence of congestion externality. Transportation Research Part A 39 (1) (2005), 17–40.

[10] Wong K.I., Bell M.G.H.: The optimal dispatching of taxis under congestion: a rolling horizon approach. Journal of Advanced Transportation, (2006).

[11] Geotaxi.

http://www.geotaxi.com/

[12] Weather Underground.

http://www.wunderground.com/

[13] Sapo Mapas.

http://mapas.sapo.pt/

[14] Mitchell, T.M.: Machine Learning. McGraw-Hill, New York, (1997).

[15] Phithakkitnukoon, S. and Dantu, R.: CPL: Enhancing Mobile Phone Functionality by Call Predicted List. The 3rd International Workshop on MObile and NEtworking Technologies for social applications (MONET'08), pp. 571-581 (2008)