Urban Mobility Study using Taxi Traces

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

Santi Phithakkitnukoon
Culture Lab, School of Computing Science, Newcastle University, Newcastle, 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, Coimbra, Portugal
bento@dei.uc.pt

ABSTRACT
In this work, we analyze taxi-GPS traces collected in Lisbon, Portugal. We perform an exploratory analysis to visualize the spatiotemporal variation of taxi services; explore the relationships between pick-up and drop-off locations; and analyze the behavior in downtime (between the previous drop-off and the following pick-up). We also carry out the analysis of predictability of taxi trips for the next pick-up area type given history of taxi flow in time and space.

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

ACM Classification Keywords
I5.2. Patter Recognition: Pattern analysis.

General Terms
Algorithms.

INTRODUCTION
In the last decades, urban areas are struggling with their growth in population and size. That demands for more resources specially energy and transportation. To maintain a constant flow of people and vehicles, we need to reduce the use of individual means of transport (e.g. car), and stimulate the use of public transportation modes (e.g. bus, metro, train). However, we need to improve the public transportation system in order to meet citizens' needs. A more efficient public transportation system can lead to a reduction of traffic congestions and consequent reduction of energy consumption. However, to optimize the public transportation network it is essential to understand what drives the common citizen, what their needs are.

At same time, we are experiencing new developments in ubiquitous computing technologies. Nowadays we are able to access to a wider variety of devices, with a growing number of features and computational capabilities. This technological diversity provides us the tools to sense urban spaces. We can either take a snapshot of all environment or follow a single vehicle or individual.

Retrieving data from the traditional public transportation (e.g. bus, train, metro) can provide a relevant database of samples and general passengers' movement. However, it does not provide the exact origin and destination for each passenger, since these transportation modes relies 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 drop-off them 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 relevant pick-up and drop-off locations in time and space; study the relationships between pick-up and drop-off locations and characterize the scenario between taxi services (i.e. what happens between the latest drop-off and next pick-up). For the latter, we explore the possibility of predicting the next pick-up area type given the previous drop-off hour of the day, day of the week, weather condition, and area type.

The paper is structured as follows: section 2 introduces the related work on urban mobility using taxi traces. Section 3 describes the source dataset, along with the environment under study. Section 4 presents a spatiotemporal study,

UbiComp' 11, Sep 17 – Sep 21, 2011, Beijing, China.

describes how taxi-GPS traces are distributed in time and space, and the relations between pick-ups and drop-offs locations. Section 5 analyzes the predictability of taxi trips for the next pick-up area type given history of taxi flow in time and space based on naïve Bayesian classifier. The final section concludes and discusses some future research to improve the current work.

RELATED WORK

The study of taxi-GPS traces to understand and improve urban mobility is a quite active research field. In this section we present some of the related work and their main contributions.

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.

DATASET AND ENVIRONMENT DESCRIPTION

In this section, we will describe the source dataset and characterize the environment under study.

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 (the realistic longest trips from one side of the city to the other could be around 22km). Data was collected from 217 distinct taxis, which account for nearly 15% of taxis in Lisbon area.

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 (red), 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 (yellow), 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 grid with cells of 0.5x0.5 km^2.

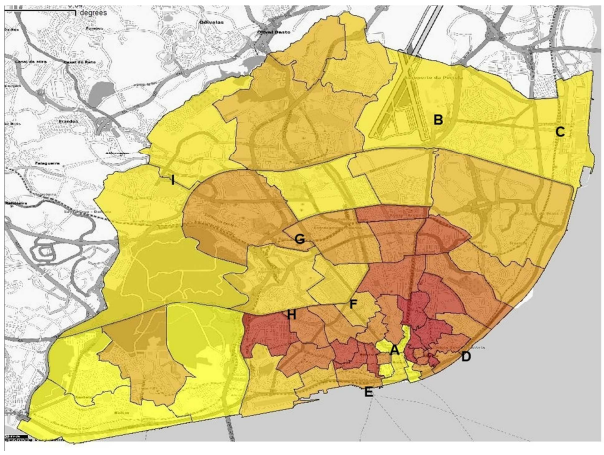


Figure 1. Lisbon council and population density (A, City downtown; B, Airport; C, Train Station; D, Train Station; E, Ferry dock; F, City center; G, Univ. Campus; H, Commercial Area; I, Residential).

Weather conditions for the period under study were retrieved from Weather Underground [11] and grouped in three states (sunny, cloudy and rainy).

Sapo Maps [13] provided a collection of 10,954 Points Of Interest (POIs), grouped into eight categories[1] (Services,

[1] The classification was performed by the data provider.

Recreation, Education, Shopping, Police, Health facilities, Transportation and Accommodation, represented in figure 2), to characterize the area type. Education facilities (e.g. kindergarten, high school, university, etc.), Recreation (e.g. bar, restaurant, etc.) and Services (e.g. bank, etc.) are the dominant POI categories (which account for over 70%).

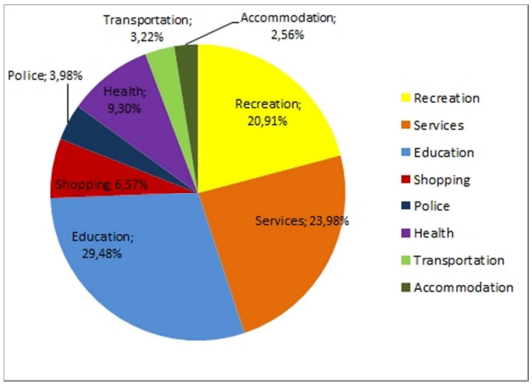


Figure 2. POIs categories distribution.

In figure 3 we can observe the raw map of POIs and the underlying density distribution. As expected the POIs are mainly distributed in areas with a higher population density or commercial. The main cluster is located in city center and downtown.

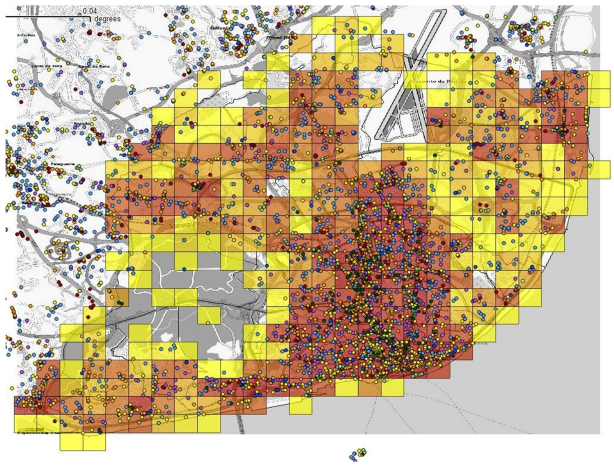


Figure 3. POI's raw map and density distribution.

Figure 4 aggregates the POI distribution in order to identify the predominant POI on each cell grid, according to figure 2 classification. On marginal street of Tagus river recreation is the most predominant POI. City center is characterized by services while education is predominant on the remaining areas.

EXPLORATORY ANALYSIS

In this section we perform an exploratory analysis to identify emerging patterns and obtain a better understanding of the variables that model the system. We explore the following aspects: spatiotemporal analysis, spatial relationships between pick-up and drop-off locations and analysis of the movement of taxis between services (i.e. from the previous drop-off to the following pick-up).

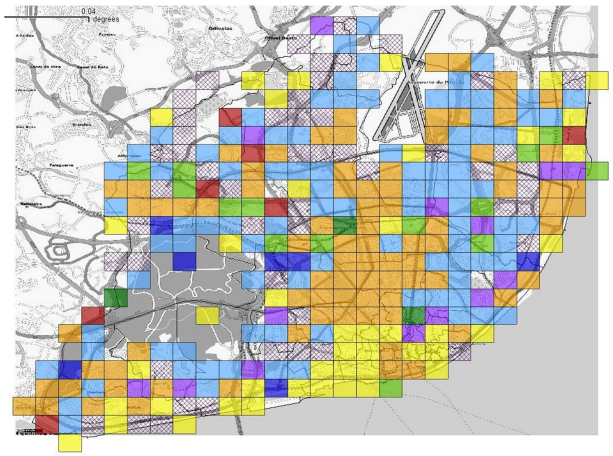


Figure 4. Predominant POI category on each location (colors correspond to classification performed in figure 2).

Spatial and Temporal distribution

Taxi demands vary in time and space, according to the citizens need. Figure 5 presents the taxi service variation according to the hours of the day and days of week.

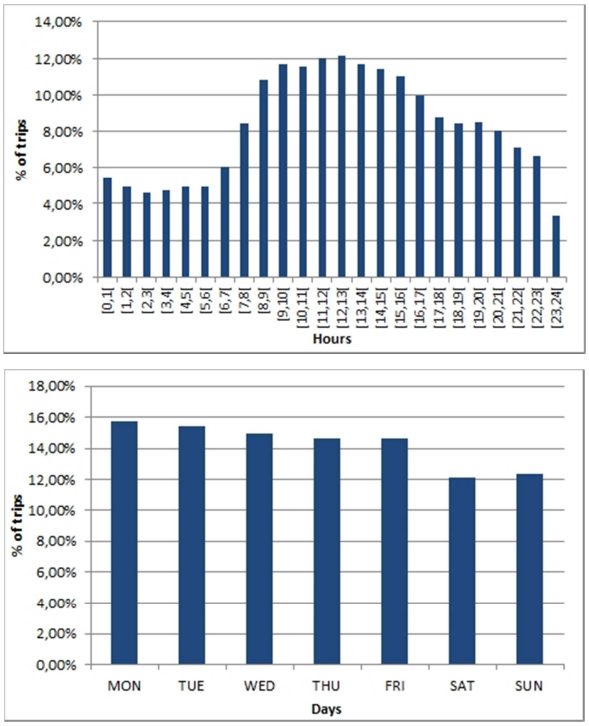
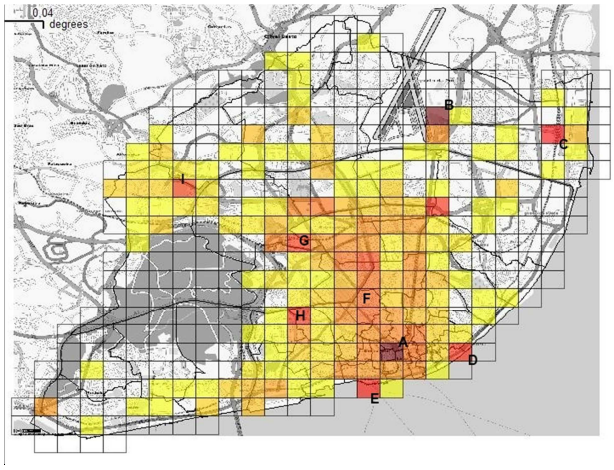


Figure 5. Taxi service variation according to the hours of day (top) and days of week (bottom)

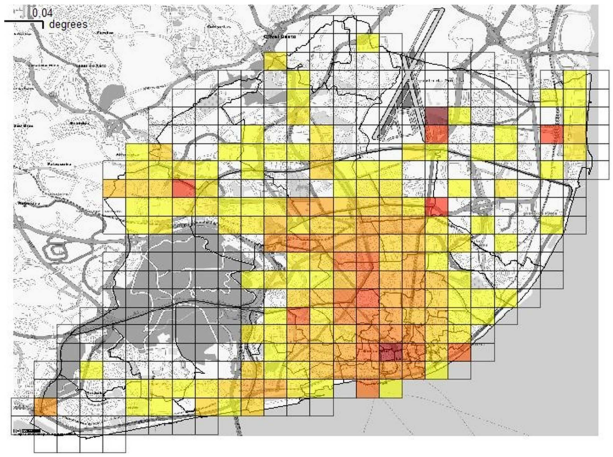
As expected, the taxi service variation follows the business hours. It gradually increases in from 7 a.m., reaches the maximum between 11 a.m. and 1 p.m., and slowly drops down in the afternoon. By the same token, there are more

taxi services in working days than in weekends. In both cases the maximum is reached in the beginning of the periods (11.am. to 1 p.m. for hours and Monday for days).

Figure 6 presents the taxi service distribution in Lisbon, according to the pick-ups and drop-offs, 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 6. 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 figure 7 we can visualize how the pick-up and drop-off location areas relate, where the thickness of the line represents the intensity between every two possible locations. 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 (airport, train, ferry, bus). 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. It is important to stress out that, although there is a subway service in Lisbon, do not exists a direct subway line connection the aforementioned locations.

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.

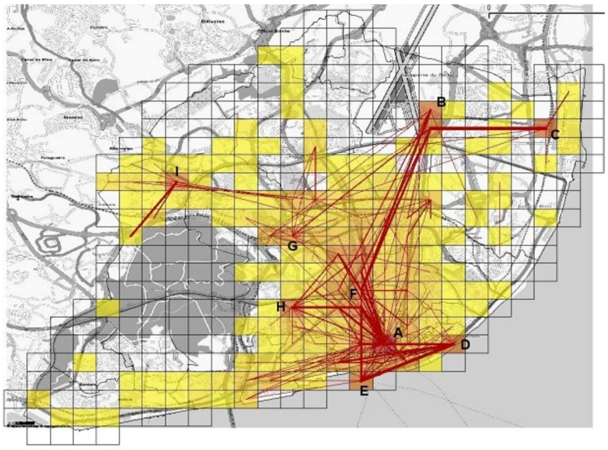


Figure 7. How strongly connected locations are, according to taxi services (A, City downtown; B, Airport; C, Train Station; D, Train Station; E, Ferry dock; F, City center; G, Univ. Campus; H, Commercial Area; I, Residential).

In figure 8 we can observe the relation between pick-ups and drop-off locations, considering only the most frequent destination for each location. By filtering the remaining destinations, we can visualize the predominant relations between locations, and their strength. Become visible the links B-C (airport and train station); A-D (downtown and train station), D-E (train station and ferry dock) and A-F (downtown and city center). Once again, a bridge between transportation modalities is observable.

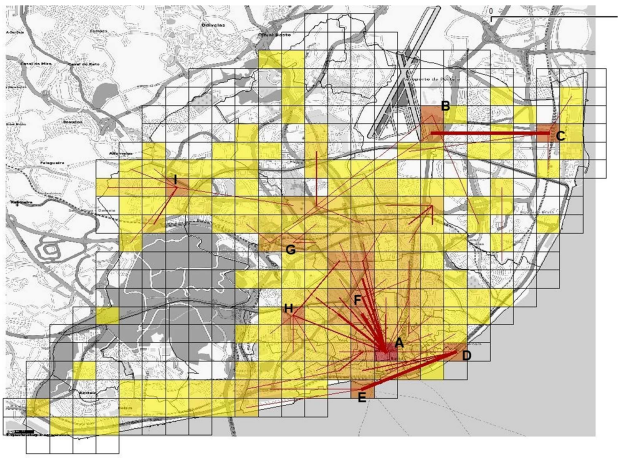


Figure 8. Relation between pick-ups and drop-off considering only the most frequent destination for each location.

To better understand the patterns from the taxi services we plot the taxi trips according to the distance, duration and income in figure 9.

[7] fitted the trips distance with a gamma distribution (with $\alpha = 2.7$ and $\beta = 1.2$). This observation does not agree with the results from different authors, where an exponential fit was observed using data collected in Florence urban area, Italy [14]. However, [7] demonstrated that exponential distribution is a special case of gamma distribution, and if the first step of the dataset is removed the trips distance could be fitted with an exponential distribution (with $\lambda = 0.26$). By the same token, if the first step of trips duration is removed, the trips duration can be fit with an exponential distribution. For trips income[2] it is not clear the fitted distribution. [1], using data collected in Shenzhen, South China observed a normal distribution for trips income.

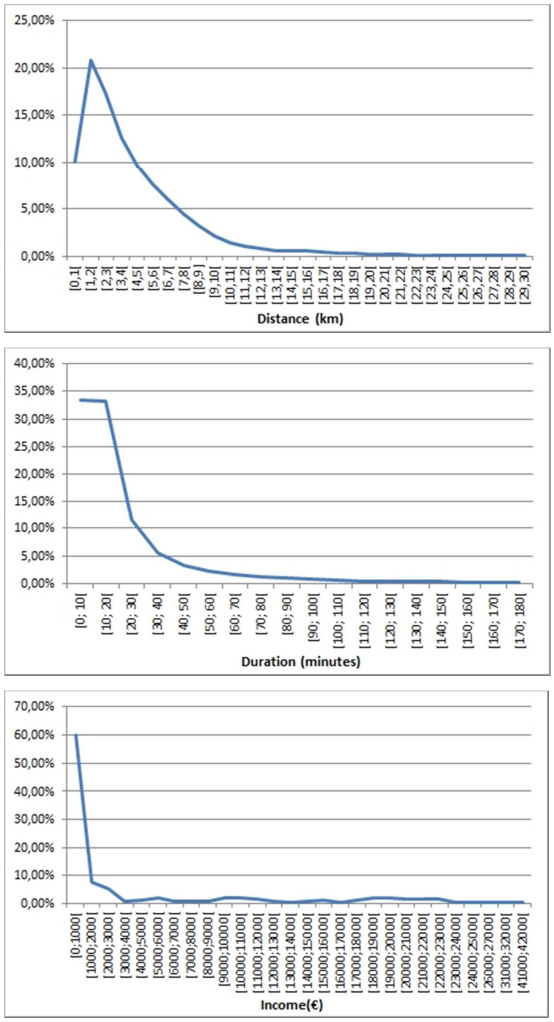


Figure 9. Taxi service distribution according to distance (top), duration (middle) and income (bottom).

The difference in results for other authors can be due the following aspects: a) distinct dataset (e.g. [1] worked with 3,000 distinct taxi drivers, whilst our dataset contains only 217 distinct taxi drivers), and b) to specific taxi drivers' behaviors (e.g. it was observed a considerable amount of trips from the airport to a nearby bus stop, and returning, locate at less than 500m, a behavior that affect the overall distributions).

[2] The income was calculated from data using the ANTRAL standard formulation http://www.antral.pt/simulador.asp. ANTRAL is a national association for transportation.

Downtime analysis

The previous analysis focused on the taxi service, in other words, the relation between the pick-up and the corresponding drop-off. Also interesting to understand is the analysis of what happens in between services (i.e. downtime – time spent looking for next pick-up), since it can help improve the taxi drivers' income.

Figure 10 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 (line thickness represents strength).

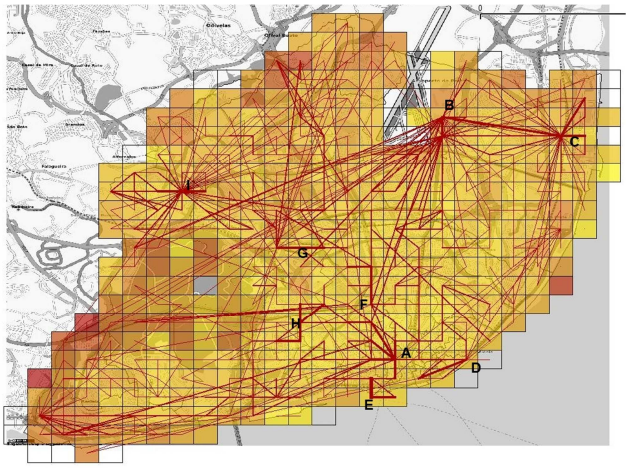


Figure 10. Spatial distribution according to the average distance traveled during downtime and the relationship between previous drop-off and next pick-up location (A, City downtown; B, Airport; C, Train Station; D, Train Station; E, Ferry dock; F, City center; G, Univ. Campus; H, Commercial Area; I, Residential).

The areas away from the city center (characterized by a higher number of residential buildings) 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.

In figure 11 we can see a density grid, where next pick-up location takes place in the same location as the previous drop-off. Downtown (A), city center (F), airport (B) and

train stations (C and D) are the locations with higher probability to pick-up a new customer in the same area after the previous drop-off.

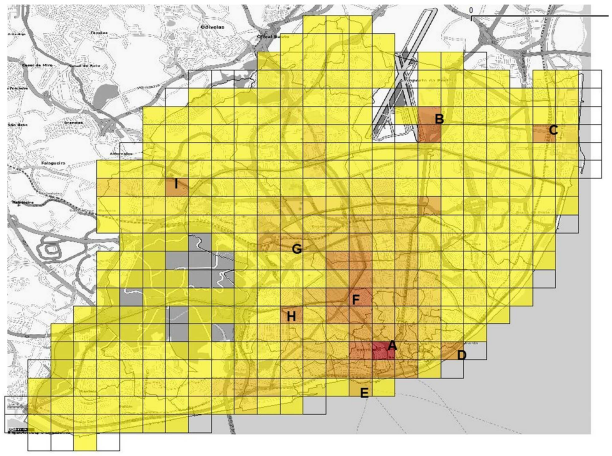


Figure 11. Next pick-up location takes place in the same location as the previous drop-off.

From these preliminary results we can estipulate that taxi drivers may want to improve their income by targeting the above-mentioned locations, or at least move to those locations after the latest drop-off, since it can improve the probability to pick-up a new customer in a reasonable amount of time and without the need to travel great distances.

Figure 12 shows the variation in trips made by and the number of taxis in service throughout the day, whereas Figure 13 shows the average time spent and distance traveled during downtime.

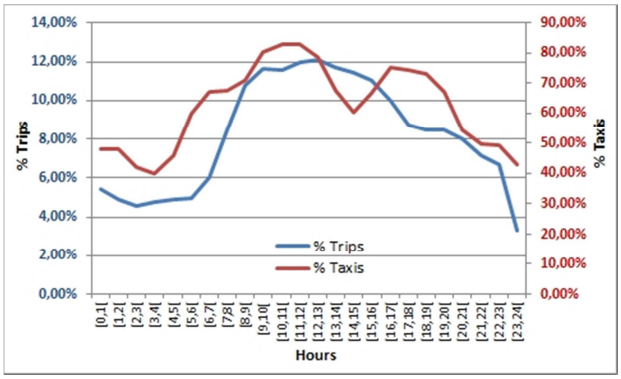


Figure 12. Amount of trips made by (blue) and number of taxis in service (red) throughout the day.

Due to the low amount of taxis in service in the early AM hours (12 a.m. to 7 a.m.), 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.

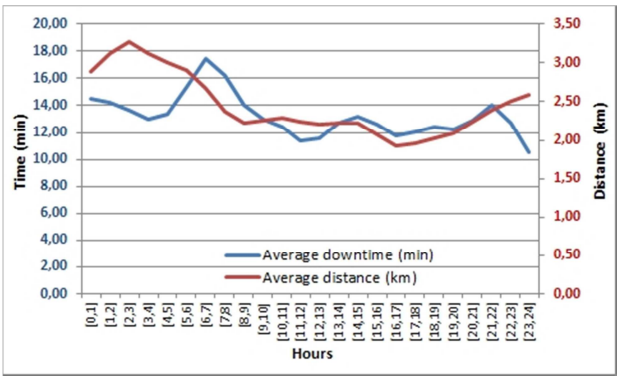


Figure 13. Average downtime (blue) and distance traveled (red).

Both distance traveled and downtime appear to follow exponential distributions as argue by [14] (figure 14).

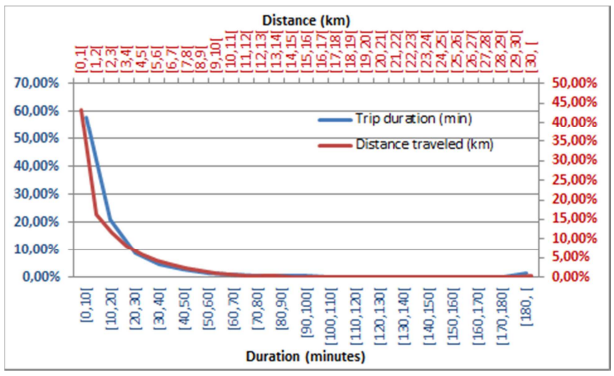


Figure 14. Distribution of distance traveled (red) and time spent (blue) during downtime.

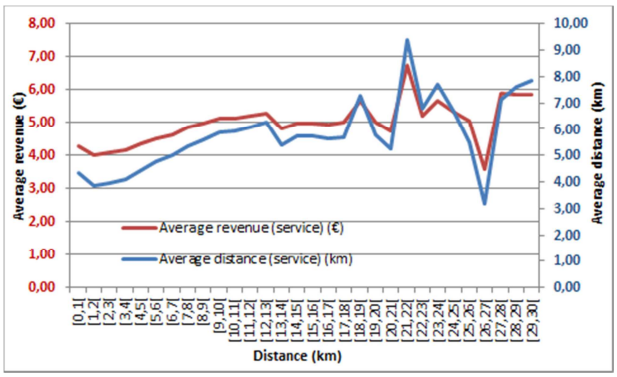


Figure 15. Variation in service distance (blue) and income (red).

In figure 15, 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.

We can conclude that, 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 (airport, train stations, ferry dock or main bus stops), 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.

PREDICTABILITY ANALYSIS

One of the main features of taxi services is the ability to adapt to the passenger needs, since it is not bounded to pre-defined path or pick-ups and drop-offs locations. Therefore, taxi movement dynamically adapts to the flow and the need of the city. The natural question would be: is it possible to predict the taxi movements? One can argue that due the apparent randomness of taxi that goal can be challenging. However, our exploratory study shows the possibility of some movement patterns (e.g. temporal and spatial density of pick-ups and drop-off, the relation between pick-ups and drop-offs).

In previous work [6] 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. 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 [15]. The objective is to compute the likelihood of each possible pick-up area type (Y) given the hour of the day (T), day of the week (D), weather condition (W) and area type (I) of the last drop-off. 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 10-folds cross validation, we are able to predict (for each drop-off) the next pick-up area type at about 54%. Previous experiments shows that individual taxi trips are relatively random and therefore a challenging problem.

CONCLUSIONS

Our work is focused in the analysis of taxi-GPS traces to better understand the urban mobility. Using traces collected in Lisbon, Portugal we are able to visualize the spatiotemporal variation, identifying the main pick-up and drop-off locations and busy hours. We also to identify the link between pick-up and drop-off locations, observing strong links between public transportation terminals, where taxi service appears to be a bridge between different public transportation services. We analyze the behavior during downtime – time spent searching for next pick-ups - where taxis tend to avoid making long trips to suburban areas for pick-up.

Our predictability analysis explores the possibility to predict the next pick-up area type given the drop-off features. With Bayesian approach given time of the day, day of the week, weather condition and area type of the current drop-off location, 54% of all trips are predictable, showing that individual taxi trips are relatively random.

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, and driving strategies to improve the income.

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. Bazzani, A., Giorgini, B., Rambaldi, S., Gallotti, R., Giovannini, L., Statistical Laws in Urban Mobility from microscopic GPS data in the area of Florence. Journal of Statistical Mechanics: Theory and Experiment, Volume 2010, (2010).
15. Mitchell, T.M.: Machine Learning. McGraw-Hill, New York, (1997)