Exploratory Study of Urban Flow using Taxi Traces

Marco Veloso[1,2], Santi Phithakkitnukoon[3,4], Carlos Bento[1], Patrick Olivier[3], Nuno Fonseca[1,2]

[1] Centro de Informática e Sistemas da Universidade de Coimbra, Portugal
[2] Escola Superior de Tecnologia e Gestão de Oliveira do Hospital, Portugal
[3] Culture Lab, School of Computing Science, Newcastle University, United Kingdom
[4] SENSE*able* City Lab, Massachusetts Institute of Technology, Cambridge, MA, USA
{mveloso, bento, nuno}@dei.uc.pt, santi@mit.edu

Abstract. The analysis of vehicle's GPS traces such as taxis can help better understand urban mobility and flow. In this paper we present a spatiotemporal analysis of taxis GPS traces collected in Lisbon, Portugal during the course of five month. We also show that trip distance can be represented with a Gamma distribution, and discuss the taxi driving strategies and respective income.

Keywords: Urban mobility, taxi traces, spatiotemporal analysis, Global Positioning System (GPS).

1 Introduction

Nowadays we experience a fast growth of the urban areas. This increase of people living and working in urban areas demands for a new public transportation system that meets their needs. With the concerns for safety and energy efficiency, there is a need for an optimized public network. To understand the urban mobility, we can make use of analysis of different transportation modes (e.g. bus, metro, taxi, private cars) through pervasive technologies (e.g. GPS, GSM). More and more these services are getting instrumented with sensors and devices that can sense a great amount of valuable data such as location and occupancy. Urban mobility can then be studied with these traces.

Taxi traces can help better understand the urban mobility as it contains the origins and destinations of the passengers with greater accuracy compared with other public transportations e.g. bus, metro. It is a unique transportation service in a way that it is not bounded to a pre-determined path or pick-up and drop-off locations. On the other hand, it is also important to consider the reasons that people use taxis instead of other public transportations. The reasons could possibly be the lack of public transportation alternatives, the urgency of the trip, the amount of luggages, or even simple commodity.

As the first step in this on-going work to better understand the urban mobility, we analyze taxi traces to identify and characterize the attractiveness of places in the city of Lisbon, Portugal. Moreover, we examine the temporal variation of the service and its relation to the predominant pick-up and drop-off locations.

2 Related Work

Liu et al [1] classify taxi drivers into the top and ordinary drivers according to the 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 route. Yuan et al [3] propose the T-Drive system that relies in 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. Chang et al [4] 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 [5] 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 overall error rate of less than one taxi per 1x1 km^2 area. There are also studies performed by Yang et al [6] and Wong et al [7] in order to improve the taxi service in congestion scenarios.

3 Data Description

Our dataset contains over than 10 million taxi-GPS samples from August to December 2009, collected in Lisbon by GeoTaxi [8]. The data sampling rate varies according to the trip nature. Samples can be stored according to the distance covered by the vehicle, time elapsed or when some state changed (e.g. occupancy). For the study purposes, only the pick-up and drop-off location and timestamp are considered, corresponding to 271,172 distinct trips. The data was collected from 253 distinct taxis, which account for nearly 15% of taxis in Lisbon area.

The area of study encompasses the Lisbon council, consisting of 53 parishes, an area around 110 km^2, with a population of 800,000. The urban area growth in several layers around the city downtown, the central point, 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). In the marginal avenue around Tagus River, there are touristic, recreational and commercial areas. Moving away from the city center we find greater parishes with lower population density, characterized by residential areas around 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.

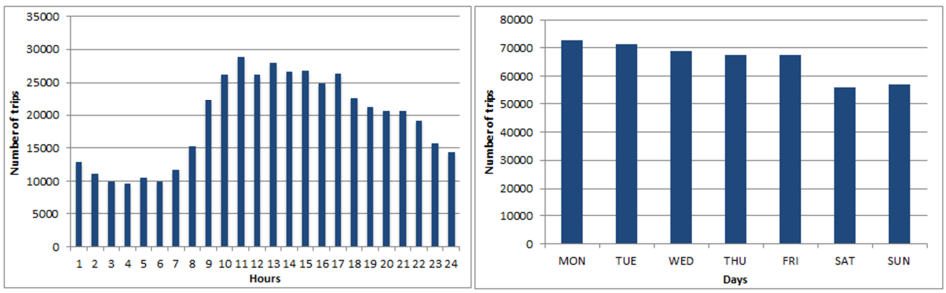
3.1 Spatial and temporal distribution

The overall taxi service distribution in Lisbon can be seen in Fig. 1 where some major locations such as the city downtown (A), the airport (B), and the train station (C) are identified.

(a) Pick-up locations (b) Drop-off locations

Fig. 1. Taxi pick-up (a) and drop-off locations (b) distribution.

The following figure shows the temporal variation of the taxi services. As expected, it gradually increases in the morning reaches the maximum between 11 a.m. and 1 p.m, and slowly drops down in the afternoon. In the working days, there are more taxi services than in the weekends with the maximum number of the services observed on Monday.



(a) Hourly variation (b) Daily variation

Fig. 2. Taxi service variation according to the hours of day (a) and days of week (b).

3.2 Data cleaning

Here we describe our data cleaning process. Based on our original dataset, a distribution of the trip distances is shown in Fig. 3. We notice that that the realistic longest trips could be around 22km (one side of the city to the other), we thus discarded trips with distance greater than 30km. On the other hand, the original data also contains a great amount of trips with less than 200m (14.94% of the all trips),

which seems unrealistic. Therefore, in addition, we discarded these small trips from our analysis.

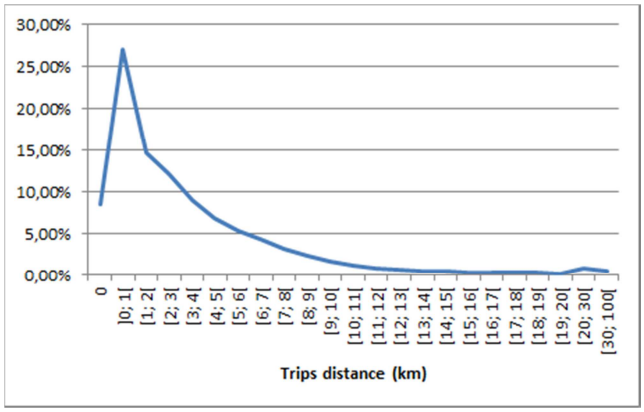


Fig. 3. Distribution of trips distances.

The trip duration distribution is represented by the following figure. Similar to the trip distance distribution, we discarded trips that are less than one minute and longer than two hours. After the data cleaning process, we retained 177,169 trips and 217 distinct taxis.

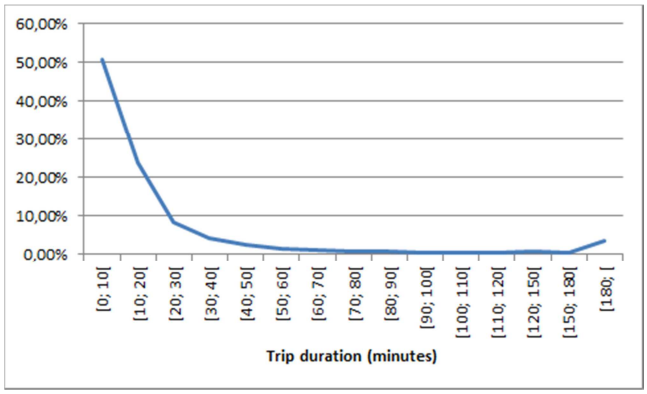


Fig. 4. Distribution of trips duration.

4 Data analysis

In the data analysis process we seek to identify the data distribution according to the trips distance, and possible driving strategies.

4.1 Trip distance distribution

After the data cleaning process, we examine the trip distance distribution (Fig. 5) and find that we can fit it with a Gamma distribution with α = 2.7 and β = 1.2 as following:

$$f_{\alpha,\beta}(x) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-\frac{x}{\beta}} \quad (1)$$

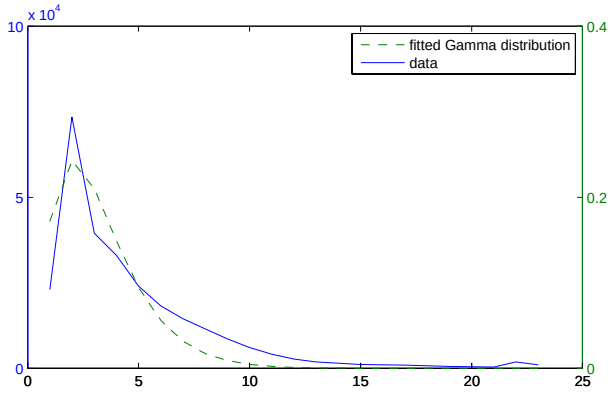


Fig. 5. Distribution of trips distance (solid line) and fitted Gamma distribution (dashed line).

If the first interval is removed, we can represent it with an exponential distribution, exp(λ) with λ = 0.26 (Fig. 6). The same phenomenon (exponential distribution) can also be observed for the trip duration and the trip income.

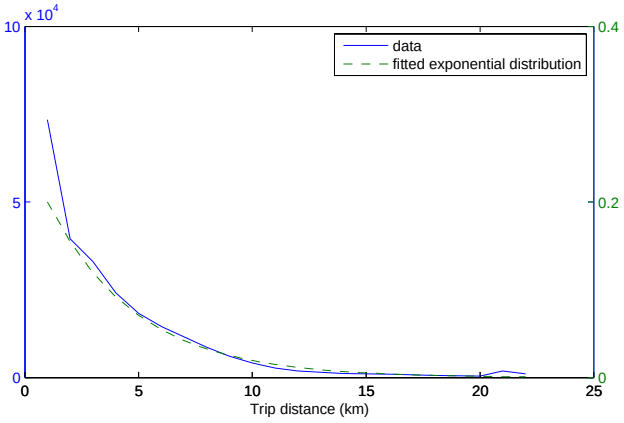


Fig. 6. Trip distance distribution without the first interval of the dataset (solid line) and fitted exponential distribution with λ = 0.26 (dashed line).

4.2 Driver strategies

We further analyze taxi driver strategies with respect to the income. A taxi driver may choose to pick up customers at a particular location (e.g. airport), or drive around the city to find passengers at random places, or combine the two approaches. We find that only 2.37% of the taxis chose to stay at the same location for more than 50% of the time. The airport seems to be one of the main pick-up and drop-off locations. Table 1 shows the statistics of the five best drivers according to the total number of trips and income (with an extra last entry for a comparison with Table 2 in a later discussion). We observe low percentages of trips from the airport from these top drivers. Table 2 shows the five best drivers according to the number of trips from the airport. These show that the taxi driver can improve the revenue by adopting a strategy of driving around the city instead of targeting one particular place like the airport.

Table 1. Statistics for the top five taxi drivers, considering total number of trips and income.

ID	Total number of trips	Total income (€)	Trips from airport	% Trips from airport	Income[1] from airport (€)
792	15,789	41,691.10 €	73	0.46 %	181.58 €
782	9,202	26,399.60 €	40	0.43 %	157.81 €
754	8,504	31,778.50 €	122	1.43 %	486.16 €
90	6,693	24,103.20 €	4	0.06 %	20.42 €
808	6,649	19,769.40 €	169	2.54 %	564.74 €
…	…	…	…	…	…
714	3,030	11,930.70 €	71	2,34 %	315.38 €

Table 2. Statistics for the top five taxi drivers, considering the number of trips and income from the airport.

ID	Total number of trips	Trips from airport	% of trips from airport	Income from airport pickup (€)	Income from other pick-up locations (€)
37	3,003	2,343	78.02 %	5,066.81 €	1,519.02 €
538	1,947	817	41.96 %	3,094.13 €	4,617.33 €
134	4,377	813	18.57 %	3,873.69 €	14,637.90 €
193	2,829	348	12.30 %	1,676.02 €	9,159.32 €
112	831	317	38.15 %	688.92 €	1,159.85 €

We find that the majority of the taxi drivers appear to be using combined strategies – mainly driving around the city and in certain time periods staying at a fixed location. This phenomenon can be observed in the Fig. 7 where there is an increase of trips from the airport between 6 a.m. and 8 a.m. and elsewhere in other time slots. On the other hand, in the same period there is a significant decrease of trips from the remaining locations.

To confirm our observation, we examine mobile phone data collected from GSM networks. The GSM data consists of samples from November and December 2009

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

providing statistical measures of carried load (Erlang) with one-hour period. In Fig. 8, we can observe that immediately after the taxi rush hour there is an increase of the GSM network usage in the airport, while the mobile phone activities in other locations of the city begins to arise three hours later. This is an indication of the possible taxi passengers at airport.

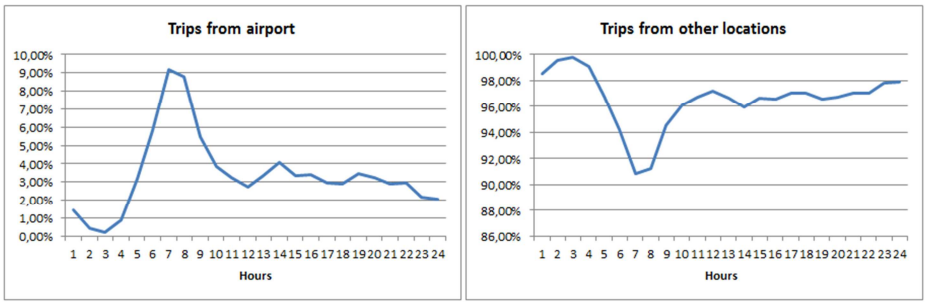


Fig. 7. Variation of number of trips aggregated by hour. Left: pick-up location from the airport. Right: pick-up from other locations.

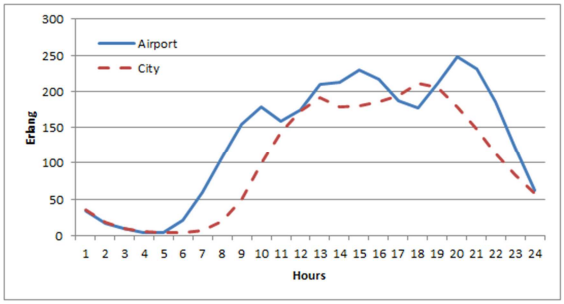
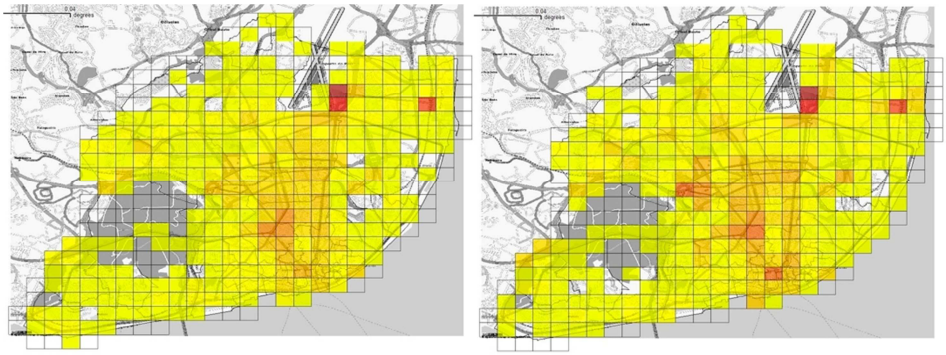


Fig. 8. GSM network usage aggregated by hour.



(a) Pick-up locations (b) Drop-off locations

Fig. 9. Distribution of (a) pick-up locations with destination the airport and (b) drop-off locations from the airport.

Another interesting observation in this preliminary study is that the main drop-off locations from the airport are (as expected) the city downtown and the main train station, but also the airport itself. The main pick-up locations with the airport being the final destination are also the city downtown, the main train station, and also (surprisingly) the airport itself. By analyzing these airport trips individually, we observe that passengers take taxis to reach a nearby bus station and parking area. These airport-to-airport trips could also be for transporting some heavy luggage between terminals.

5 Conclusions

Analyzing the taxi traces from Lisbon, Portugal during a period of five months, we are able to capture the spatiotemporal variation and observe that trip distance, duration, and income follow Gamma and Exponential distribution. We also examine taxi driving strategies where a combination of pick-up customers around the city and in fixed locations is observed. There are clearly rooms for improvements and further investigations for this preliminary study. As our future work, we will continue to investigate on how taxi traces can be useful for urban planning and transportation management. We will also explore into data visualization and multi-source data fusion – integrating taxi data with other public transportation such as bus, metro, and fleets as well as mobile phone data to better understand the mobility and flow of the city.

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 pp. 105-120 (2010)
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: Proceedings of the 10th international conference on Ubiquitous computing, New York, NY, USA, ACM 322-331 (2008)
3. Yuan, J., Zheng, Y., Zhang, C., Xie, W., Xie, X., Huang, Y.:T-Drive: Driving Directions Based on Taxi Trajectories, in ACM SIGSPATIAL GIS 2010, Association for Computing Machinery, Inc., 1 (2010)
4. Chang, H., Tai, Y., Hsu, J.Y.: Context-aware taxi demand hotspots prediction. Int. J. Bus. Intell. Data Min. 5(1) pp. 3-18 (2010)
5. Phithakkitnukoon, S., Veloso, M., Bento, C., Biderman, A., Ratti, C.: Taxi-Aware Map: Identifying and predicting vacant taxis in the city. In First International Joint Conference on Ambient Intelligence (2010)
6. 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): pp. 17–40 (2005)
7. Wong K.I., Bell M.G.H.: The optimal dispatching of taxis under congestion: a rolling horizon approach. Journal of Advanced Transportation (2006)
8. Geotaxi, http://www.geotaxi.com