Taxi Driver Assistant – A Proposal for a Recommendation System

MARCO VELOSO
Centre for Informatics and Systems of the University of Coimbra, Portugal
College of Management and Technology of Oliveira do Hospital, Portugal
and
SANTI PHITHAKKITNUKOON
Open University, United Kingdom
and
CARLOS BENTO
Centre for Informatics and Systems of the University of Coimbra, Portugal

Taxi is a flexible way of transportation; however, with the fast growth of urban areas, the task of finding new passenger quickly becomes a challenging task for taxi drivers. In this work, we propose a recommendation system to assist taxi drivers searching for the next pick-up. With an inference engine based on a naïve Bayesian classifier, we analyze taxi-GPS traces collected in Lisbon, Portugal. An exploratory and predictive analysis is performed to better understand the apparent randomness of taxi movements, and the impact of features' variation, namely, temporal periods, driver strategies or cell sizes, on the classifier.

Categories and Subject Descriptors: I.5.2. Patter Recognition: Pattern analysis
General Terms: Algorithms
Additional Key Words and Phrases: Urban mobility, taxi-GPS traces, spatiotemporal analysis, predictive analysis, naïve Bayesian classifier, recommendation system

1. INTRODUCTION

The evolution of the society led to several changes in the organization of the current demography. With a fast growth, rapidly the urban areas are supporting the majority of the population. Among other demands, there is the need to maintain a constant flow of people and vehicles. However, to optimize the public transportation network it is essential to understand what drives the common citizen.

Authors' addresses: Marco Veloso (email: mveloso@dei.uc.pt), Centre for Informatics and Systems of the University of Coimbra, Portugal; College of Management and Technology of Oliveira do Hospital, Portugal; Santi Phithakkitnukoon (e-mail: santi@newcastle.ac.uk), Open University, United Kingdom; Carlos Bento (email: bento@dei.uc.pt), Centre for Informatics and Systems of the University of Coimbra, Portugal.

Taxi service is a flexible way of transportation, since it is not bounded to pre-defined path or pick-ups and drop-offs locations. Taxi movement dynamically adapts to the flow and the need of the city: it can pick-up the passengers right where they are standing, and drop-off them precisely in the desirable destination. Therefore, taxi service can provide a more accurate information about the origins and destinations of passengers, when compared to other traditional public transportation modes (e.g. bus, metro, train). Nevertheless, with the growth of urban areas, it becomes more difficult no move within the cities, and to be efficient in the searching for new passengers.

At the same time, we are experiencing new developments in ubiquitous computing technologies. A wide variety of devices is available with an increasing processing and storage capability, along with a diversity of functions and sensors. These devices have been used widely to collect data from the taxi movement and, therefore, study their patterns in search for improvements.

We envision a system that could help the taxi drivers, by making recommendations of locations likely to have potential passengers. These recommendations are supported by an inference engine and a database of past paths of the current driver and the taxi community. The system should be flexible to allow the taxi driver to select the desirable features that should be taken into account by the inference engine, namely, current location, hour of the day, day of the week, weather conditions, or proximity of points of interest (POI).

Our on-going work is focused on the analysis of taxi-GPS traces acquired in the city of Lisbon, Portugal, to propose and test the recommendation system, currently in development. The contribution of this work lies on the following aspects:

1. a spatiotemporal analysis of a dataset of taxi-GPS traces,
2. a proposal for a recommendation system and its inference engine,
3. a study of the predictability of taxi volume and its sensibility to variations of features.

For the former, we analyze taxi traces to identify relevant pick-up and drop-off locations in time and space; study the relationships between those locations. 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. For the second contribution we propose a simple recommendation system based on a naïve Bayesian classifier, and explore the impact on the predictability of features' variation, namely, temporal window, taxi driver behavior or size of the destination locations.

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, describes how taxi-GPS traces are distributed in time and space, and the relations between pick-ups and drop-offs locations. Section 5 describes the recommendation system and the inference engine based on naïve Bayesian classifier. A study of the effects of different features in the predictability is performed analyzing differential temporal windows, taxi driver behaviors, cell types and cell sizes.

2. RELATED WORK

Taxi-GSP traces have been used in a number of studies to develop better solutions and services in urban areas such as estimating optimal driving paths [Yuan 2010, Zheng 2010, Ziebart 2008], predicting next taxi pick-up locations [Yuan 2011, Chang 2010, Phithakkitnukoon 2010, Liu 2010a, Ge 2010], modeling driving strategies to improve taxi's profit [Ge 2010, Liu 2010b], identifying flaws and possible improvements in urban planning [Zheng 2011], and developing models for urban mobility, social functions, and dynamics between the different city's areas [Qi 2011, Veloso 2011].

Yuan et al. [2010] present the T-Drive system that identifies optimal route for a given destination and departure time. Zheng et al. [2010] describe a three-layer architecture using the landmark graph to model knowledge of taxi drivers. Ziebart et al. [2008] present a decision-modeling framework for probabilistic reasoning from observed context-sensitive actions. The model is able to make decisions regarding intersections, route, and destination prediction given partially traveled routes.

Yuan et al. [2011] develop a recommender system for both taxi drivers and passengers that takes into account the passengers' mobility patterns and taxi drivers' pick-up traces. Chang et al. [2010] propose a four-step approach for mining historical data in order to predict taxi demand distributions based on time, weather, and taxi location. They show that different clustering methods have different performances on distinct data distributions. Phithakkitnukoon et al. [2010] present a model for predicting the number of vacant taxis for a given area of the city based on the naïve Bayesian classier with their developed error-based learning algorithm and a mechanism for detecting adequacy of historical data. Liu et al. [2010a] classify taxi drivers according to their income. They observe that top drivers operate in a number of different zones while maintaining exceptional balance between taxi demand and traffic conditions. Ordinary drivers on the other hand operate in fixed zones with few variations.

Ge et al. [2010] present an approach for extracting energy-efficient transportation patterns from taxi traces and use it to develop a recommender system for pick-up locations and a sequence of waiting locations for a taxi driver. Zheng et al. [2010b] identify flawed urban planning in region pairs with traffic problems and the linking structure among these regions through their analysis of taxi traces. Qi et al. [2011] investigates the relationship between regional pick-up and drop-off characteristics of taxis and social function of city regions. They develop a simple classification method to recognize regions' social areas that can be divided into scenic spots, entertainment districts, and train/coach stations. Veloso et al. [2011] present an exploratory analysis of the spatiotemporal distribution of taxi pick-ups and drop-offs. They investigate downtime (time spent looking for next passengers) behavior, identify taxi-driving strategies, and explore relationship between area type (based on points of interest) and taxi flow, as well as the predictability of a taxi trip.

3. DATASETS

This work analyzes taxi volume in Lisbon, Portugal. The data was collected from August through December in 2009.

3.1 The city

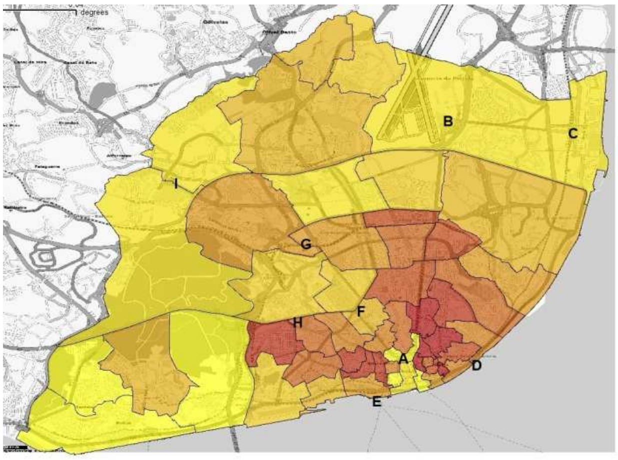

Fig. 1. Lisbon council and population density (red means more density). A, City downtown; B, Airport; C, Train Station; D, Train Station; E, Ferry dock; F, City center; G, Univ. Campus; H, Commercial Area; I, Residential.

The area of study encompasses the Lisbon council (Figure 1) that consists of 53 parishes, an area of around 110 km2, 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 500x500m2-grid cells (397 cells on total).

Weather conditions for the period under study were retrieved from Weather Underground[1] and grouped in three states (sunny, cloudy and rainy).

3.2 Taxi

Our taxi dataset was provided by GeoTaxi[2], a company that focuses on software development for fleet management, and holds about 20% of the taxi market share in Portugal. The dataset was composed of around 10 million taxi-GPS location points and collected from 217 taxis. Along with the GPS location (latitude, longitude) information, it reported speed, bearing, engine status, and occupancy status of the taxi. For study purposes, only pick-up and drop-off locations and timestamps are considered, which correspond to 177,169 distinct trips (Figure 2). 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), and less than a minute and longer than three hours.

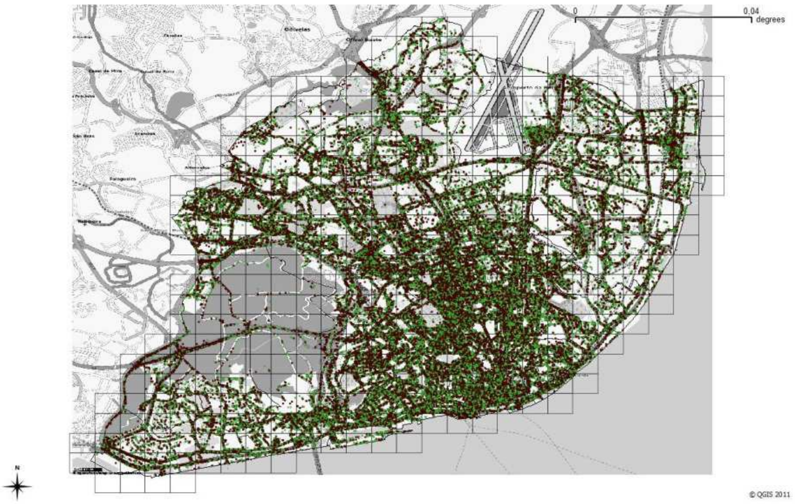

Fig. 2. Spatial distribution of pick-ups (red) and drop-offs (green).

[1] http://www.wunderground.com/

[2] http://www.geotaxi.com/

3.3 Points of Interest

Sapo Maps[3] provided a collection of 10,954 points of interest, grouped into eight categories (Services, Recreation, Education, Shopping, Police, Health facilities, Transportation and Accommodation, represented in Figure 3), 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%).

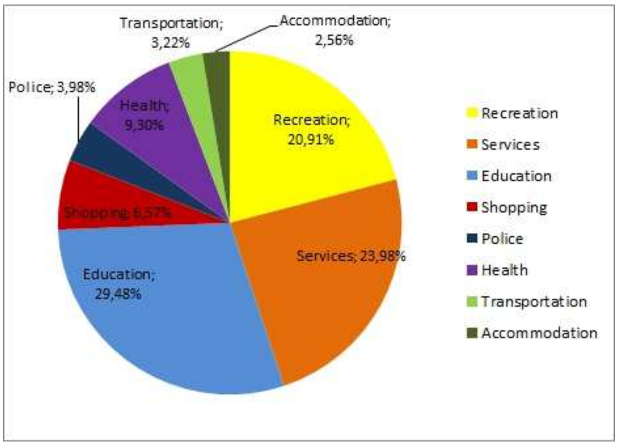

Fig. 3. POIs categories distribution.

In Figure 4 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.

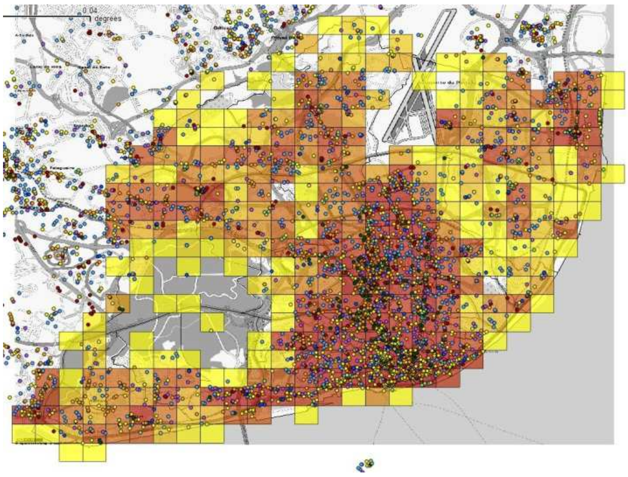

Fig. 4. POI's raw map and density distribution.

[3] http://mapas.sapo.pt/

Figure 4 aggregates the POI distribution in order to identify the predominant POI on each cell grid, according to Figure 3 classification. Near the Tagus River shore recreation is the most predominant POI. City center is characterized by services while education is predominant on the remaining areas.

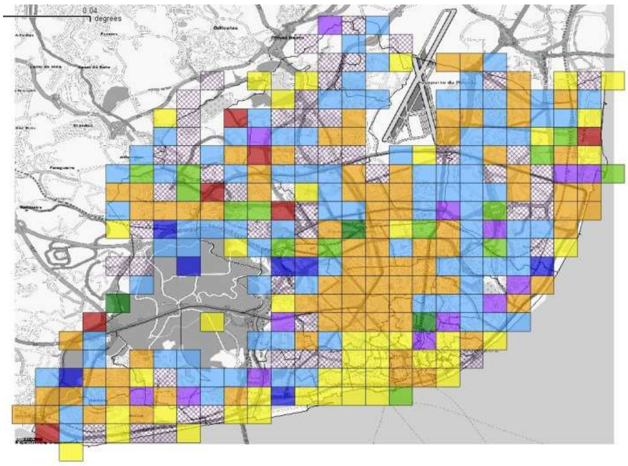

Fig. 5. Predominant POI category on each location (colors correspond to classification performed in Figure 2).

4. EXPLORATORY ANALYSIS

Taxi volume varies in time and space, according to the citizens need. Figure 6 presents the taxi service variation according to the hours of the day and days of week.

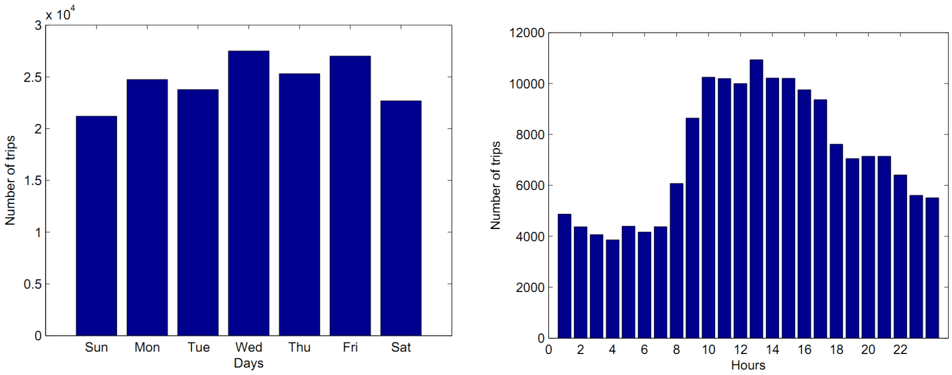

Fig. 6. Taxi service variation according to the and days of week (left) and hours of day (rigth).

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).

The overall taxi volume’s spatial distribution in Lisbon is shown in Figure 7 (on 500x500m2-grid cells), where the number of pick-ups on each cell during the period under study is represented by a color scale (red corresponds to cells with a higher number of pick-ups). Some major locations are identified, such as city downtown (A), airport (B), train stations (C, D) and ferry dock (E). Different public transportation modalities (airport, train, ferry, bus) are well connected through taxi services.

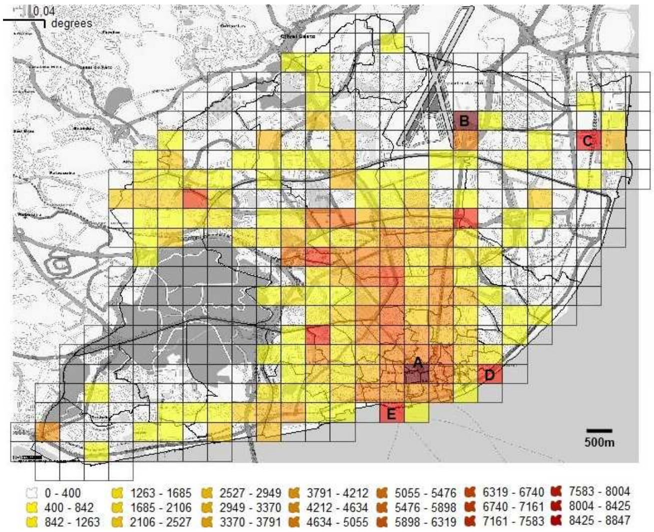

Fig. 7. Spatial distribution of taxi volume (number of pick-ups).

In figure 8 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.

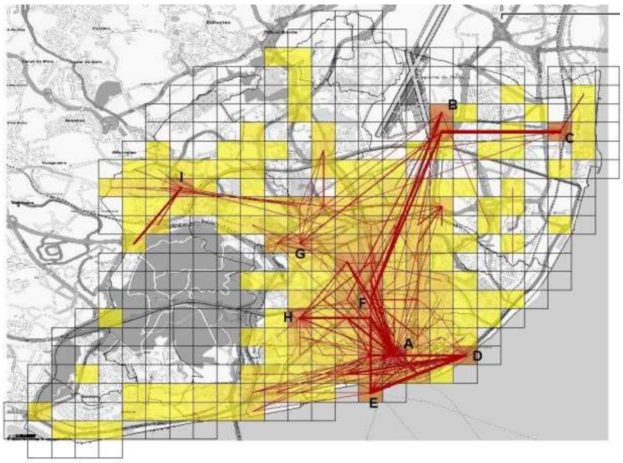

Fig. 8. How strongly connected locations are, according to taxi services (A, City downtown; B, Airport; C, Train Station; D, Train Station; E, Ferry dock; F, City center; G, Univ. Campus; H, Commercial Area; I, Residential).

To better understand the patterns from the taxi services we plot the taxi trips according to the distance, duration and income in Figure 8.

On our previous work-in-process conference paper [2011], of which this paper is an extension, we were able to fit the trips distance with a gamma distribution (with $\alpha = 2.7$ and $\beta = 1.2$). This observation does not agree with the results from other authors, where an exponential fit was observed using data collected in Florence urban area, Italy [Bazzani 2010]. However, the former demonstrated that exponential distribution is a special case of gamma distribution, and if the first steps of the dataset are removed the trips distance could be fitted with an exponential distribution (with $\lambda = 0.26$). By the same token, if the first steps of trips duration are removed, the trips duration can be fit with an exponential distribution. For trips income[4] it is not clear the fitted distribution. [Liu 2010], using data collected in Shenzhen, South China observed a normal distribution for trips income.

We believe that small values of trip duration, trip distance and trip income are in fact noisy data. Realistically, a taxi trip must be longer than 200m or one minute (minimal values accepted during the cleaning process). Since we were unable to prove that believe, we decided to be conservative during the cleaning process. If these low values were removed (corresponding to the first steps of the distribution), the trip duration, distance

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

and income would follow an exponential distribution, as predicted by Bazzani et al. [2010].

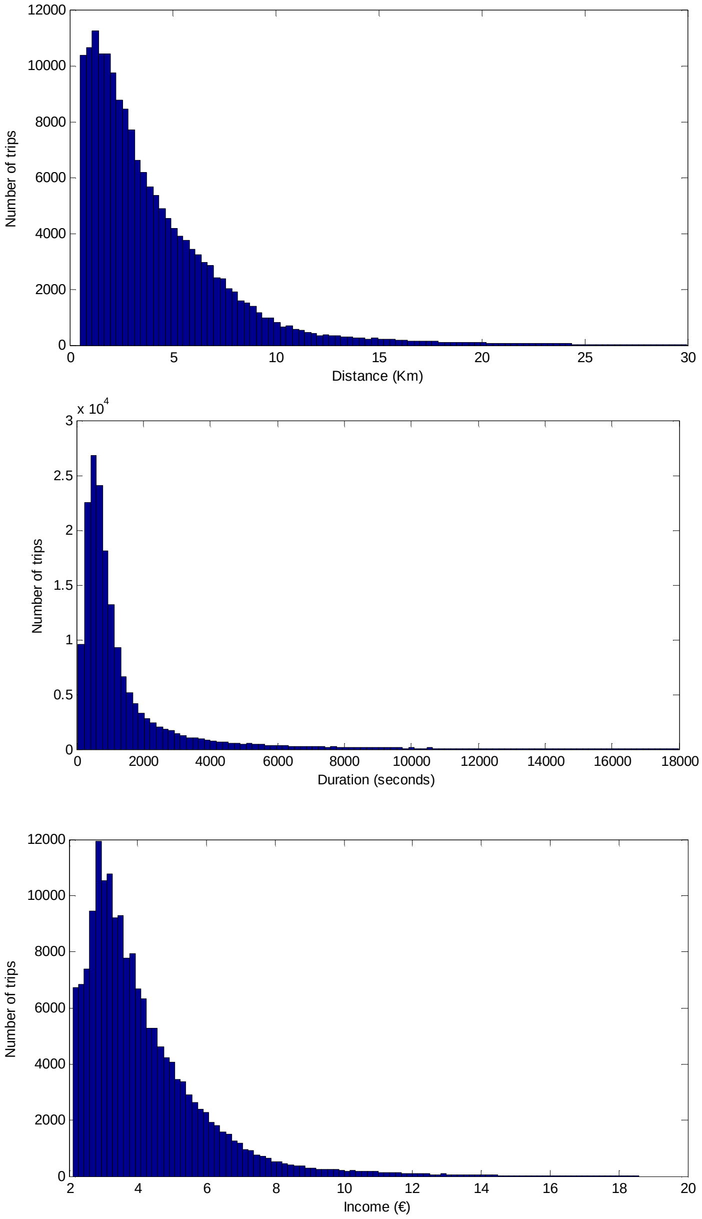

Fig. 9. Taxi service distribution according to distance (top), duration (middle) and income (bottom).

The difference in results for other authors can be due the following aspects: a) distinct dataset (e.g. Liu et al. [2010] worked with 3,000 distinct taxi drivers, whilst our dataset contains only 217 distinct taxi drivers), 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), and c) due possible noisy data.

5. TAXI DRIVER RECOMMENDATION SYSTEM

Taxi service is a flexible way of transportation, 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. This flexibility can led to an apparent randomness and the prediction of taxi movements can be challenging. However, day of the week, time of the day, and weather condition are promising features in predicting taxi volume and 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).

Our goal is to develop a recommender system, which could help the taxi drivers to decide what the next pick-up location would be. Based on a Naïve Bayesian Classifier, the system should provide the likelihood of findings passengers on the urban space. Because the area is modeled by a flexible grid system, the user can interact with the system and obtain a personalized visualization.

5.1 System Overview

A basic system overview is shown in Figure 9 (left). In order to make a recommendation, the system extracts data from a database of taxi-GPS traces, and pre-processes it to select relevant features that match the current scenario, namely, location, day of the week, hour, weather condition or area type (characterized by POI). The selected data is processed by a classifier and the output presented to the user (the taxi driver).

The graphical interface is formed by three layers that provide enriched information, and allows the user to filter or select specific views (Figure 10 right). The first layer represents the map of the urban area. The second layer provides the likelihood of each possible pick-up location according to the available data from the taxi community. The third layer provides a similar information but using only the historic data from the current driver. This configuration should be flexible to allow the user to select the desirable features that should be taken into account by the inference engine, and provide controls so the user could zoom in and zoom out to explore with detail the map with the likelihood information.

The data used for the inference engine could contain samples from the taxi community or only past information from the current driver. By the same token, the system could provide a personalized recommendation, considering only the current location (samples were the previous drop-off match the current location), or a global information, taking in account all history available (samples were the previous drop-off took place at any location). Additionally, besides the current location, the system could also process information from neighbors cells to improve the accuracy of the system.

The likelihood information is presented within a grid and a scheme of colors (e.g. red more likely to find a passenger, yellow, less likely). The size of the cells can be changed to meet the user demand. However, as it is demonstrated in the next section, the use of smaller cells will reduce the performance of the inference engine.

After each drop-off, the new taxi-GPS trace should be cleaned and transformed to be stored in the database.

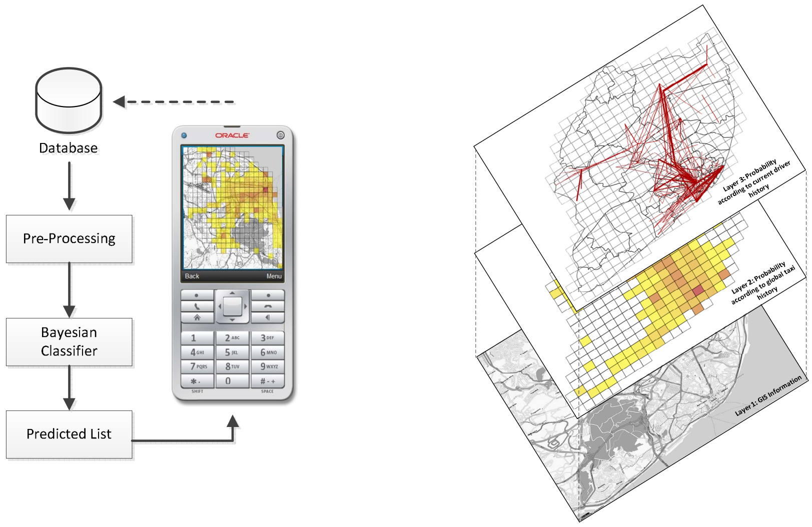
Database
Pre-Processing
Bayesian Classifier
Predicted List
ORACLE
Back
Menu
Layer 3: Probability according to current driver history
Layer 2: Probability according to global taxi history
Layer 1: GIS information

Fig. 10. Basic system overview (left) and system's visualization layers (right).

5.2 Inference Engine

Our inference engine is based on the Naïve Bayesian Classifier, which is a simple probabilistic classifier based on Bayes' theorem with independence assumptions. Bayes rule of conditional probability [Mitchell 1997] is defined by:

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}, \qquad (1)$$

where *P(A|B)* is the posterior probability, which is the probability *A* given the feature *B*, *P(B|A)* is the likelihood of *A* with respect to *B*, *P(A)* is called prior probability and *P(B)* the evidence factor.

In this work, we want to compute the likelihood of each possible pick-up location (*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:

$$P(Y = y_i | T, D, W, I) = \frac{P(Y = y_i)P(T, D, W, I | Y = y_i)}{P(T, D, W, I)}, \qquad (2)$$

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:

$$\begin{aligned} y_{MAP} &= \arg\max_{y_i \in Y} P(Y = y_i | T, D, W, I) \\ &= \arg\max_{y_i \in Y} P(Y = y_i)P(T, D, W, I | Y = y_i) \\ &= \arg\max_{y_i \in Y} P(Y = y_i) \prod_i P(T | Y = y_i)P(D | Y = y_i)\, P(W | Y = y_i)P(I | Y = y_i), \end{aligned} \qquad (3)$$

4.2 Results analysis

Global accuracy

Based on 10-folds cross validation we are able to predict the next pick-up location, at about 7%. Similar values were obtained using different training sets (e.g. 5-folds cross validation, hold out 2/3 versus 1/3). To understand the effect of the features hour of the day, day of the week, weather condition and area type, the classifier was modified to take in consideration only the current location (previous drop-off). The next pick-up location was correctly predicted at about 5%. This reducing, although small, is an indication of the positive effect of the aforementioned features in the classification process.

To further investigate the low predictability aspect, we examine the predicted list [Phithakkitnukoon 2008] – 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 11 (left) shows that accuracy rate varying with the length of the predicted list. The list must grow to about 70 possible destinations in order to predict the correct destination at a high accuracy (70%). This can reflect the randomness of the taxi flow in the city, the fact the variables could not be independent, and the wide search space.

Figure 11 (right) shows the evolution of the accuracy with the number of testing samples classified, for a hold out 2/3 versus 1/3 (around 49000 samples). The accuracy value of at about 5% stables after 17% of the testing samples being classified.

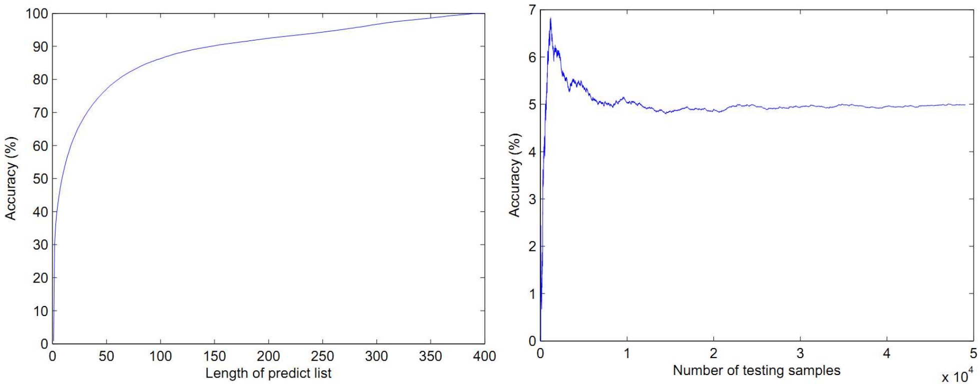

Fig. 11. Corresponding accuracy rate for growing length of the predicted list (left) and accuracy rate for number of testing cells (right).

Effect of daily and weekly periods

From the exploratory spatiotemporal analysis we observed temporal patterns where active hours of the day (8AM-8PM) and active days of the week (weekdays) present a slightly higher predictability (Figure 12). This behavior can be explained by the existence of more activities (mostly repeated activities in temporal orders such as commuting to work, lunch time at same similar place, school activities, and so on) on weekdays and active hours than on weekend. A similar observation was performed on previous work were mobile phone call activity was had a strong correlation with taxi volume on weekdays and active hours.

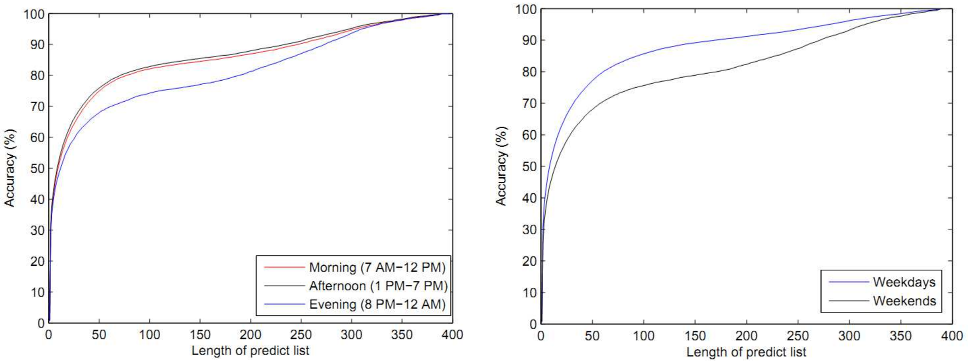

Fig. 12. Overall performance of the predicted list for different daily periods (left) and between weekdays and weekends (right).

Effect of taxi driver strategies

Taxi drivers have different strategies to identify the best location for the next pick-up or a fastest path to drop-off. Liu et al. [2010a, 2010b] and Zheng et al. [2010, 2011] explored this behavior to model knowledge of taxi drivers. Liu et al. [2010a] classified drivers as top drivers and ordinary drivers according to the income. In our work we also explore the difference between top drivers, average drivers and low performance drivers, considering their income and amount of trips. Although top drivers present specific strategies (e.g. searching for passengers near the airport between 7 AM and 9 AM, when the majority of international flights from the West arrive to Lisbon Airport), they are also characterized for high amount of trips achieved.

Figure 13 shows the overall performance of the predicted list for top, average and low performance drivers. The former rapidly increases the accuracy of prediction, needing to grow to 37 possible destinations in order to predict the correct destination at a high accuracy (70%). Average drivers must grow to 177 and low performance drives to about 250 to attain similar accuracy. Since average and low performance drivers do not appear to have specific driving strategies to improve their income, their behavior can be characterized by certain randomness.

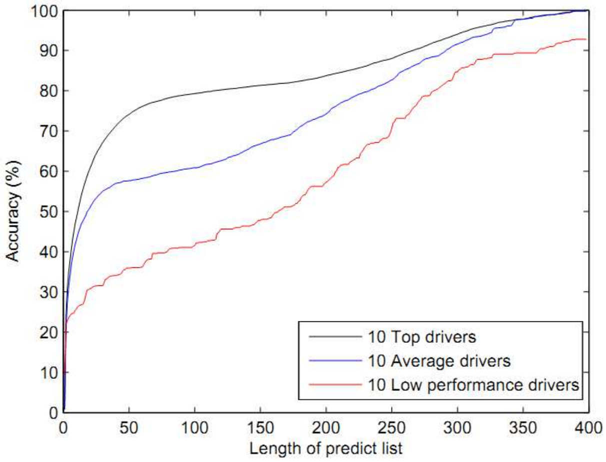

Fig. 13. Overall performance of the predicted list for different taxi drivers' type (top, average and low performance drivers).

Randomness or uncertainty associated with a random variable has been studied and defined as the information entropy by Claude E. Shannon [Shannon 1948] as follows:

$$E(X) = -\sum_{i} p(x_i) \log_2 p(x_i) \, , \qquad (4)$$

where *E(X)* is an entropy of random variable *X* where $x_i \in X$ and $p(x_i) = Prob(X = x_i)$. Similar to [Phithakkitnukoon 2008] we use the information entropy to define the randomness of the finding the next pick-up (*X*). Unsurprisingly, top drivers have a lower value of entropy (4.8842) than the average driver (4.9650). Moreover, since the low performance drivers are characterized by a low amount of trips, the scarce amount of training data could also influence the performance of the classifier.

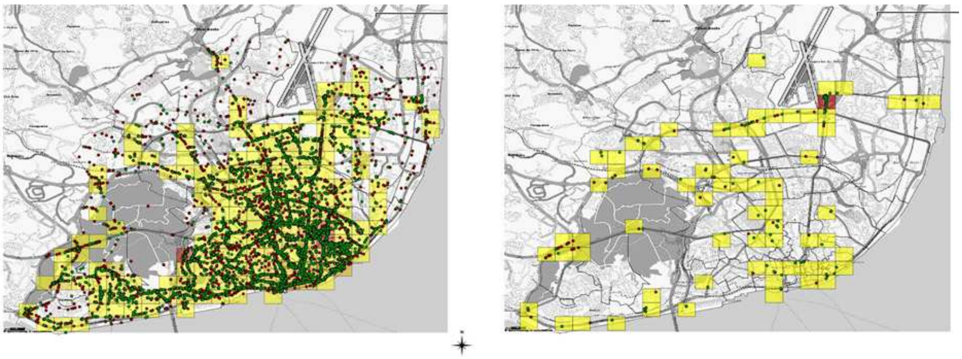

Fig. 14. Pick-ups and drop-offs distribution for a top driver (left) and a low activity driver (right).

Figure 14 shows the pick-ups and drop-offs for a top driver and a low activity driver. Although the top driver has a high amount of pick-ups and drop-offs sparse throughout the city, several clusters are visible, which can help the prediction process. The low activity driver has a single predominant location, but with several isolated pick-ups and drop-offs. Usually, average and low performance drivers choose to stay at same pre-defined location waiting for the next pick-up, hence the usual single predominant location. The consequence is longer waiting times. On the other hand, top drivers are more active and change locations through the day, reducing the waiting period. This constant movement should have impact on the predictability. However, if we consider hour of the day we can observe a pattern (e.g. airport between 7 AM and 9 AM).

To improve the system efficiency, the data from low performance drivers should be removed.

Effect of cell type

The spatial analysis demonstrated that different locations have different attractability for taxi movements. For instance, previous studies have shown that taxi service is often used as a bridge between transportation modalities, such as train stations, airports or ferry docks. Therefore, those locations have a higher activity of taxi volume. We divided the cells in three groups according to the taxi activity: high (or predominant cells), normal, and low taxi activity. Figure 15 (left) shows the accuracy rate varying with the length of the predicted list for the three cells' types.

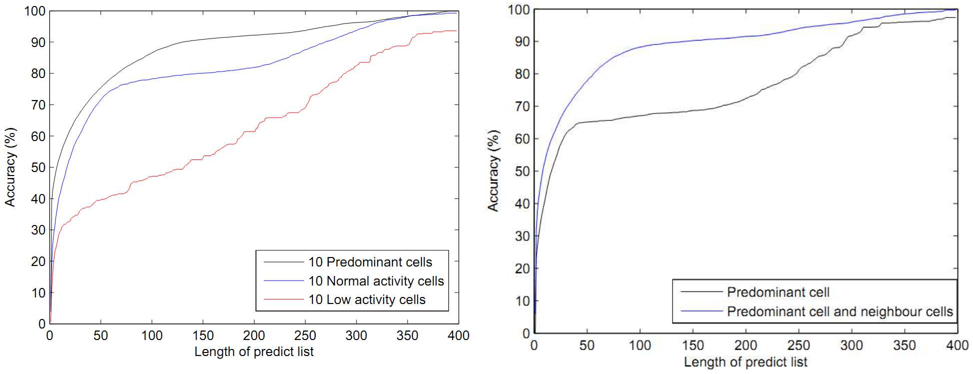

Fig. 15. Overall performance of the predicted list for high, normal, and low activity cells (left), and the contribution of neighbor cells (right).

The predictability of taxi activity on predominant cells rapidly reaches a high accuracy (70%) with the first 37 possible destinations. Normal activity cells quickly attain the same value with the first 47 possible destinations, while low activity cells need at about 250 possible destinations from the prediction list. The corresponding entropy values are: 4.3851, 4.7395 and 4.8141. Similar to the taxi driver analysis, the scarce amount of training data could also influence the performance of the classifier for the cells with low taxi activity. We also observe a better accuracy predicting the next pick-up location, when the data from the neighbor cells of the current location are also included in the classification process (Figure 15 right).

Effect of cell size

One important feature of the recommendation system is to allow the user to zoom in and out the interface to obtain a more detailed information of a specific location (Figure 16). However, when the cell size is reduced the performance of the inference engine diminishes (Figure 17). By reducing the size of each cell, there is an increase of the number of possible destinations and a substantial reduction of the amount of instances on each cell.

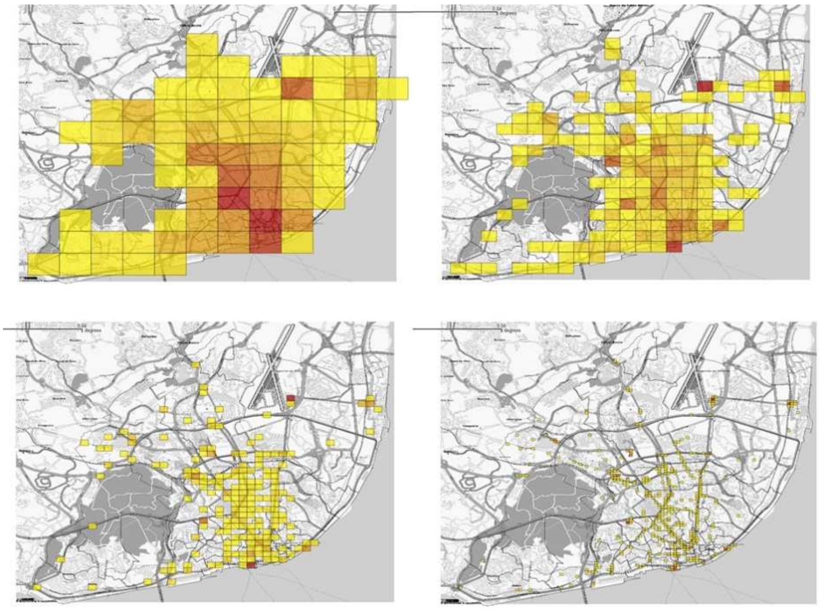

Fig. 16. Spatial distribution of taxi volume for different cell size: 1000mx1000m (top left), 500mx500m (top right), 250mx250m (bottom left) and 100mx100m (bottom right). Cells with the least amount of trips were removed.

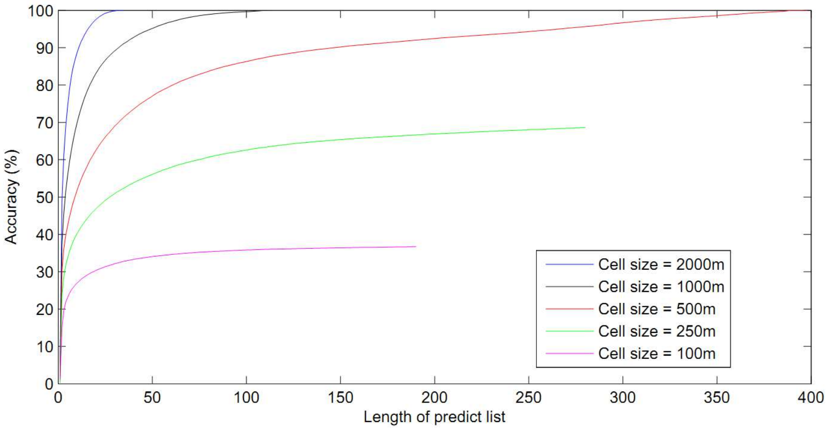

Fig. 17. Overall performance of the predicted list for different cell sizes.

6. CONCLUSIONS

Taxi service is a flexible way of transportation, and dynamically adapts to the flow and the need of the city. However, the fast growth of urban areas difficult the process of searching for new passenger efficiently (reducing the waiting time and the distance traveled to the next pick-up).

In this work we propose a recommendations system, currently in development, to assist the taxi driver in the task of picking-up new passengers. The system is based on a naïve

Bayesian classifier and is characterized by the ability to use different sets of features (computing the likelihood according to location, day of the week, hour, weather condition and area type) and to adapt to the user needs (use different cell sizes and data from the current location, neighbor cells, or all locations available).

Our analysis shows that taxi flow can be very random. Although top drivers follow specific strategies to improve their incoming, the less active behavior of average and low performance drivers can affect the predictability of the system. The initial low accuracy of 7% can be justified by this randomness and for the high amount of possible destinations. The prediction list must grow to about 70 possible destinations to be able to predict the correct destination with a high accuracy (70%). Several effects to the classifier were explored, namely, the influence of daily and weekly periods and the impact of cell's size and type.

Our goal is to conclude the current development and deploy and test the system on a real environment, to evaluate the system recommendations according to the driver decision. The inference engine, the initial assumptions (e.g. variables independency) and the cell's size must be revised as they have a great influence on the system's performance.

REFERENCES

BAZZANI, A., GIORGINI, B., RAMBALDI, S., GALLOTTI, R., GIOVANNINI, L. 2010. "Statistical Laws in Urban Mobility from microscopic GPS data in the area of Florence". Journal of Statistical Mechanics: Theory and Experiment, Volume 2010.

LIU, L., ANDRIS, C., BIDERMAN, A., RATTI, C. 2010a. "Uncovering cabdrivers' behavior patterns from their digital traces". In Computers, Environment and Urban Systems.

LIU, L., ANDRIS, C., BIDERMAN, A., RATTI, C. 2010b. "Revealing taxi drivers mobility intelligence through his trace". In Movement-Aware Applications for Sustainable Mobility: Technologies and Approaches, 105-120.

CHANG, H., TAI, Y., HSU, J.Y. 2010. "Context-aware taxi demand hotspots prediction". International Journal on Business Intelligence Data Mining 5(1) , 3-18.

MITCHELL, T.M. 1997. "*Machine Learning*". McGraw-Hill, New York.

PHITHAKKITNUKOON, S., VELOSO, M., BENTO, C., BIDERMAN, A., RATTI, C. 2010. "Taxi-Aware Map: Identifying and predicting vacant taxis in the city". In Proc. AmI 2010, First International Joint Conference on Ambient Intelligence, 86-95.

PHITHAKKITNUKOON, S. AND DANTU, R. 2008. "CPL: Enhancing Mobile Phone Functionality by Call Predicted List". In 3rd International Workshop on MObile and NEtworking Technologies for social applications (MONET'08), pp. 571-581.

QI, G., LI, X., LI, S., PAN, G., WANG, Z., ZHANG, D. 2011. "Measuring Social Functions of City Regions from Large-scale Taxi Behaviors". In PerCom-Workshops 2011, pp. 21-25, Seattle, USA.

SHANNON , C. E. 1948. "A mathematical theory of com-munication" Bell System Technical Journal, vol. 27, pp. 379–423 and 623–656.

SONG, C., QU. Z. BLUMM, N., BARABÁSI, A. 2010. "Limits of Predictability in Human Mobility". In Science Vol. 327 no. 5968 pp. 1018-1021.

VELOSO, M., PHITHAKKITNUKOON, S., BENTO, C. 2011. “Urban Mobility Study using Taxi Traces”. In International Workshop on Trajectory Data Mining and Analysis (TDM) in conjunction with the 13th International Conference on Ubiquitous Computing (UbiComp), ACM Digital Library and UbiComp Extended Proceedings, Beijing, China.

YUAN, J. , ZHENG, Y., ZHANG, C., XIE, W., XIE, X., HUANG, Y. 2010. ”T-Drive: Driving Directions Based on Taxi Trajectories”. in Proc. ACM SIGSPATIAL GIS 2010, Association for Computing Machinery, Inc. 1, 99-108.

YUAN, J., ZHENG, Y., ZHANG, L., XIE, X., SUN, G. 2011. “Where to Find My Next Passenger?”. In 13th ACM International Conference on Ubiquitous Computing (UbiComp 2011), China.

ZHENG, Y., LIU, Y., YUAN, J., XIE, X. 2011. “Urban Computing with Taxicabs”. In 13th ACM Int. Conference on Ubiquitous Computing (UbiComp 2011), China.

ZHENG, Y., YUAN, J., XIE, W., XIE, X., SUN, G. 2010. “Drive Smartly as a Taxi Driver”. In 7th Int. Conference on Ubiquitous Intelligence & Computing and 7th Int. Conference on Autonomic & Trusted Computing (UIC/ATC), 484-486.

ZIEBART, B.D., MAAS, A.L., DEY, A.K., BAGNELL, J.A. 2008. “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, 322-331.