On Recommending Urban Hotspots to Find Our Next Passenger

Luis Moreira-Matias [1,2,3], Ricardo Fernandes [1], João Gama[2,5], Michel Ferreira [1,4], João Mendes-Moreira [2,3],Luis Damas [6]

[1] Instituto de Telecomunicações, University of Porto, Portugal
[2] LIAAD/INESC TEC, Porto, Portugal
[3] Department of Informatics Engineering, Faculty of Engineering, University of Porto, Portugal
[4] Department of Computer Sciences, Faculty of Sciences, University of Porto, Portugal
[5] Faculty of Economics, University of Porto, Portugal
[6] GEOLINK, Porto, Portugal
{*luis.m.matias, jgama, joao.mendes.moreira*}*[at]inescporto.pt,* {*rjf, michel*}*[at]dcc.fc.up.pt, luis[at]geolink.pt*

Abstract

The rising fuel costs is disallowing random cruising strategies for passenger finding. Hereby, a recommendation model to suggest the most passenger-profitable urban area/stand is presented. This framework is able to combine the 1) underlying historical patterns on passenger demand and the 2) *current* network status to decide which is the best zone to head to in each moment. The major contribution of this work is on how to combine well-known methods for learning from data streams (such as the historical GPS traces) as an approach to solve this particular problem. The results were promising: 395.361/506.873 of the services dispatched were correctly predicted. The experiments also highlighted that a fleet equipped with such framework surpassed a fleet that is not: they experienced an average waiting time to pick-up a passenger 5% lower than its competitor.

1 Introduction

The taxis became crucial for human mobility in medium/large-sized urban areas. They provide a direct, comfortable and speedy way to move in and out of big town centers - as complement to other transportation means or as a main solution. In the past years, the city councils tried to guarantee that the running vacant taxis will always meet the demand in their urban areas by emitting more taxi licenses than the necessary. As result, the cities' cores are commonly crowded by a huge number of vacant taxis - which take *desperate measures* to find new passengers such as random cruise' strategies. These strategies have undesirable side effects like large wastes of fuel, an inefficient traffic handling, an increase of the air pollution.

The taxi driver mobility intelligence is one of the keys to mitigate this problems. The knowledge about where the services (i.e. the transport of a passenger from a pick-up to a drop-off location) will actually emerge can truly be useful to the driver – especially where there are more than one competitor operating. Recently, the major taxi fleets are equipped with GPS sensors and wireless communication devices. Typically, these vehicles will transmit information to a data center about their location and the events undergoing like the passenger pick-up and drop-off. These historical traces can reveal the underlying running mobility patterns. Multiple works in the literature have already explored this kind of data successfully with distinct applications like smart driving [Yuan *et al.*, 2010], modeling the spatiotemporal structure of taxi services [Deng and Ji, 2011; Liu *et al.*, 2009; Yue *et al.*, 2009], building passenger-finding strategies [Li *et al.*, 2011; Lee *et al.*, 2008] or even predicting the taxi location in a passenger-perspective [Phithakkitnukoon *et al.*, 2010]. Despite their useful insights, the majority of the techniques reported are offline, discarding the main advantages of this signal (i.e. a streaming one).

Figure 1: Taxi Stand choice problem.

In our work, we focus on the online choice problem about which is the best taxi stand to go to after a passenger drop-off (i.e. the stand where we will pick-up another passenger quicker). Our goal is to use the vehicular network communicational framework to improve their reliability by combining all drivers' experience. In other words, the idea is to forecast how many services will arise in each taxi stand based on the network past behavior to feed a recommendation model to calculate the best stand to head to. An illustration about our problem is presented in Fig. 1 (the five blue dots represent possible stands to head to after a passenger drop-off; our recommendation system outputs one of them as the best choice at the moment).

Such recommendation model can present a true advantage for a fleet when facing other competitors, which will work with less information than you do. This tool can improve the informed driving experience by transmitting to the driver

which is the stand where 1) he will wait less time to get a passenger in; or where 2) he will get the service with the greatest revenue.

The **smart stand-choice problem** is based on **four key decision variables**: the expected price for a service over time, the distance/cost relation with each stand, how many taxis are already waiting at each stand and the passenger demand for each stand over time. The taxi vehicular network can be a ubiquitous sensor of taxi-passenger demand from where we can continuously mine the reported variables. However, the work described here will just address the decision process based on the last three variables.

In our previous work [Moreira-Matias *et al.*, 2012], we already proposed a model to predict the spatiotemporal distribution of the taxi passenger demand (i.e. the number of services that will emerge along the taxi stand network). This study departed from this initial work to extend it along three different dimensions:

1. The **Recommendation System**: we use these predictions as input to a **Recommendation System** that also accounts the number of taxis already in a stand and the distance to it. Such framework will improve the taxi driver mobility intelligence in real time, helping him to **decide** which **is the most profitable stand in each moment**. It will be based not only in his own past decisions and outcomes, but on a combination of everyone experience, **taking full advantage of the ubiquitous characteristics of the vehicular communicational networks**.
2. **Test-bed**: Our experiments took advantage of the vehicular network **online information** to feed the predictive framework. Moreover, the recommendation performance was **evaluated in real-time**, demonstrating its **robustness** and its ability to **learn, decide and evolve without a high computational effort**;
3. **Dataset**: 506.873 services were dispatched to our 441 vehicle fleet during our experiments. This large scale test was carried out along 9 months.

There are some works in the literature related with this problem, namely: 1) mining the best passenger-finding strategies [Li *et al.*, 2011; Lee *et al.*, 2008], 2) dividing the urban area into attractive clusters based on the historical passenger demand (i.e.: city zones with distinct demand patterns) [Deng and Ji, 2011; Liu *et al.*, 2009; Yue *et al.*, 2009] and even 3) predicting the passenger demand at certain urban hotspots [Li *et al.*, 2012; Kaltenbrunner *et al.*, 2010; Chang *et al.*, 2010]. The **major contribution** of this work facing this state-of-the-art is to **build smart recommendations about the taxi stand to head to in an online streaming environment** (i.e. real-time; while the taxis are operating) based not only on their historical trace but also on the current network status. In fact, the reported works present offline frameworks and/or test-beds or just account a low number of decision variables.

The results were obtained using two distinct test-beds: firstly, (1) we let the stream run continuously between August 2011 and April 2012. The predictive model was trained during the first five months and it was stream-tested in the last four. Secondly, (2) we used a traffic simulator to test if our Recommendation System could beat the drivers' expected behavior. We simulated a competitive scenario – with two fleets - using the services historical log and on the existing road network system. The obtained results validated that our method can effectively help the drivers to decide where they can achieve more profit.

The remainder of the paper is structured as follows. Section 2 formally presents our predictive model while Section 3 details our recommendation one. The fourth section describes our case study, how we acquired and preprocessed the data used as well as some statistics about it. The fifth section describes how we tested the methodology in a concrete scenario: firstly, we introduce the two experimental setups and the metrics used to evaluate both models. Then, the obtained results are detailed, followed by some important remarks. Finally, conclusions are drawn.

2 The Predictive Model

In this section we present some relevant definitions and a brief description of the predictive model on taxi passenger demand. The reader should consult the section II in [Moreira-Matias *et al.*, 2012] for further details. Let $S = \{s_1, s_2, ..., s_N\}$ be the set of N taxi stands of interest and $D = \{d_1, d_2, ..., d_j\}$ a set of j possible passenger destinations. Our problem is to choose the best taxi stand at the instant t according with our forecast about passenger demand distribution over the time stands for the period $[t, t + P]$.

Consider $X_k = \{X_{k,0}, X_{k,1}, ..., X_{k,t}\}$ to be a discrete time series (aggregation period of P-minutes) for the number of demanded services at a taxi stand k. The goal is to build a model which determines the set of service counts $X_{k,t+1}$ for instant $t + 1$ and per taxi stand $k \in \{1, ..., N\}$. To do so, three distinct short-term prediction models are proposed, as well as a well-known data stream ensemble framework to use all models. We briefly describe these models along this section.

2.1 Time Varying Poisson Model

Consider the probability for n taxi assignments to emerge in a certain time period - $P(n)$ - following a **Poisson Distribution**.It is possible to define it using the following equation

$$P(n; \lambda) = \frac{e^{-\lambda}\lambda^n}{n!} \quad (1)$$

where λ represents the rate (average demand for taxi services) in a fixed time interval. However, in this specific problem, the rate λ is not constant but time-variant. Therefore, it was adapted as a function of time, i.e. $\lambda(t)$, transforming the Poisson distribution into a non homogeneous one. Let λ_0 be the average (i.e. expected) rate of the Poisson process over a full week. Consider $\lambda(t)$ to be defined as follows

$$\lambda(t) = \lambda_0 \delta_{d(t)} \eta_{d(t),h(t)} \quad (2)$$

where $\delta_{d(t)}$ is the relative change for the weekday $d(t)$ (e.g.: Saturdays have lower day rates than Tuesdays); $\eta_{d(t),h(t)}$ is the relative change for the period $h(t)$ in the day $d(t)$ (e.g. the peak hours); $d(t)$ represents the weekday 1=Sunday, 2=Monday, ...; and $h(t)$ represents the period when time t falls (e.g.

the time 00:31 is contained in period 2 if we consider 30-minutes periods).

2.2 Weighted Time Varying Poisson Model

The model previously presented can be faced as a time-dependent average which produces predictions based on the long-term historical data. However, it is not guaranteed that every taxi stand will have a highly regular passenger demand: actually, the demand in many stands can often be **seasonal**. The sunny beaches are a good example on the demand seasonality: the taxi demand around them will be higher on summer weekends rather than other seasons along the year.

To face this specific issue, a weighted average model is proposed based on the one presented before: the goal is to increase the relevance of the demand pattern observed in the recent week (e.g. what happened on the previous Tuesday is more relevant than what happened two or three Tuesdays ago). The weight set ω is calculated using a well-known time series approach to these type of problems: the Exponential Smoothing [Holt, 2004]. This model will enhance the importance of the mid-term historical data rather than the long-term one already proposed in the above section.

2.3 Autoregressive Integrated Moving Average Model

The two previous models assume the existence of a regular (seasonal or not) periodicity in taxi service passenger demand (i.e. the demand at one taxi stand on a regular Tuesday during a certain period will be highly similar to the demand verified during the same period on other Tuesdays). However, the demand can present distinct periodicities for different stands. The ubiquitous features of this network force us to rapidly decide if and how the model is evolving so that it is possible to adapt to these changes instantly.

The AutoRegressive Integrated Moving Average Model (ARIMA) [Box *et al.*, 1976] is a well-known methodology to both model and forecast univariate time series data such as traffic flow data [Min and Wynter, 2011], electricity price [Contreras *et al.*, 2003] and other short-term prediction problems such as the one presented here. There are two main advantages to using ARIMA when compared to other algorithms. Firstly, 1) it is versatile to represent very different types of time series: the autoregressive (AR) ones, the moving average ones (MA) and a combination of those two (ARMA); Secondly, 2) it combines the most recent samples from the series to produce a forecast and to update itself to changes in the model. A brief presentation of one of the simplest ARIMA models (for non-seasonal stationary time series) is presented below following the existing description in [Zhang, 2003] (however, our framework can also detect both seasonal and non-stationary series). For a more detailed discussion, the reader should consult a comprehensive time series forecasting text such as the one presented in Chapters 4 and 5 in [Cryer and Chan, 2008].

2.4 Sliding Window Ensemble Framework

Three distinct predictive models have been proposed which focus on learning from the long, medium and short-term historical data. However, a question remains open: Is it possible to combine them all to improve our prediction? Over the last decade, regression and classification tasks on streams attracted the community attention due to their drifting characteristics. The ensembles of such models were specifically addressed due to the challenge related to this type of data. One of the most popular models is the weighted ensemble [Wang *et al.*, 2003]. This error-based model was employed in this framework. The Averaged Weighted Error(AVE) metric was used to measure such error.

3 Recommendation Model

Let $X_{k,t+1}$ be the number of services to be demanded in the taxi stand k during the 30-minutes period next to the time instant t. Then, a passenger is dropped-off somewhere by a vehicle of interest w minutes after the last forecast on the instant$t$. The problem is to choice one of the possible taxi stands to head to. This choice is related with four key variables: the expected price for a service over time, the distance to each stand, how many taxis are already waiting at each stand and the predicted passenger demand. However, here we solve this issue like a *minimization* problem: we want to rank the stands according the minimum waiting time (target variable) to pick-up a passenger, whenever it is directly picked-up or dispatched by the central.

Let $C_{k,t+1}$ be the number of taxis already parked in the stand k in the drop-off moment and $L_{k,w}$ be the number of services departed from the same stand between this moment and the moment of the last forecast (i.e.: t).We can define the service deficit - $SD_{k,t+w}$ on the taxi stand k i.e.: a prediction on the number of services that still will be demanded in the stand discounting the vehicles already waiting in the line) as

$$SD_{k,t+w} = (X_{k,t+1} - C_{k,t+1} - L_{k,w}) * \rho_H \qquad (3)$$

where ρ_H is the similarity (i.e.: 1 – error) obtained by our forecasting model in this specific stand during the sliding training window H. In fact, ρ_H works as a *certainty* about our prediction (i.e.: if two stands have the same *SD* but our model is experiencing a bigger error in one of them, the other stand should be picked instead).

Let υ_k be the distance (in kilometres) between the drop-off location and the taxi stand k. We can define the normalized distance to the stand - U_k - as follows

$$U_k = 1 - \frac{\upsilon_k}{\xi} \qquad (4)$$

where ξ is the distance to the farthest stand. We can calculate the Recommendation Score of the taxi stand k as

$$RS_k = U_k * SD_{k,t+w} \qquad (5)$$

Then, we calculate the Recommendation Score of every stands and we recommend to the driver the stand with the highest one.

4 Data Acquisition and Preprocessing

The stream events data of a taxi company operating in the city of Porto, Portugal, was used as case study. This city is the center of a medium-sized urban area (consisting of 1.3 million inhabitants) where the passenger demand is lower than

the number of running vacant taxis, resulting in a huge competition between both companies and drivers. The data was continuously acquired using the telematics installed in each one of the 441 running vehicles of the company fleet throughout a non-stop period of nine months. This study just uses as input/output the services obtained directly at the stands or those automatically dispatched to the parked vehicles (more details in the section below). This was done because the passenger demand at each taxi stand is the main feature to aid the taxi drivers' decision.

Statistics about the period studied are presented. Table 1 details the number of taxi services demanded per daily shift and day type. Table 2 contains information about all services per taxi/driver and cruise time. The *service* column in Table 2 represents the number of services taken by the taxi drivers, while the second represents the total cruise time of every service. Additionally, it is possible to state that the central service assignment is 24% of the total service (*versus* the 76% of the service requested directly on the street) while 77% of the service is demanded directly to taxis parked in a taxi stand (and 23% is assigned while they are cruising). The average waiting time (to pick-up passengers) of a taxi parked at a taxi stand is 42 minutes while the average time for a service is only 11 minutes and 12 seconds. Such low ratio of busy/vacant time reflects the current economic crisis in Portugal and the regulators' inability to reduce the number of taxis in the city. It also highlights the importance of the predictive system presented here, where the shortness of services could be mitigated by obtaining services from the competitors.

5 Experimental Results

In this section, we firstly describe the experimental setup developed to test our predictive model on the available data. Secondly, we introduce our simulation model and the experiments associated with. Thirdly, we present our Recommendation System and the metrics used to evaluate our methods. Finally, we present the results.

5.1 Experimental Setup for the Predictive Model

Our model produces an online forecast for the taxi-passenger demand at all taxi stands at each P-minutes period. Our test-bed was based on *prequential* evaluation: data about the network events was continuously acquired.

Each data chunk was transmitted and received through a socket. The model was programmed using the R language. The prediction effort was divided into three distinct processes running on a multicore CPU (the time series for each stand is independent from the remaining ones) which reduced the computational time of each forecast. The pre-defined functions used and the values set for the models parameters are detailed along this section.

An aggregation period of 30 minutes was set (i.e. a new forecast is produced each 30 minutes; P=30) and a radius of 100 m (W = 100 ¿ 50 defined by the existing regulations). It was set based on the average waiting time at a taxi stand, i.e. a forecast horizon lower than 42 minutes.

The ARIMA model (p,d,q values and seasonality) was firstly set (and updated each 24h) by learning/detecting the underlying model (i.e. autocorrelation and partial autocorrelation analysis) running on the historical time series curve for each considered taxi stand. To do so, we used an automatic time series function in the [forecast] R package [Yeasmin and Rob, 1999] - *auto-arima* – with the default parameters. The weights/parameters for each model are specifically fit for each period/prediction using the function arima from the built-in R package [stats].

The time-varying Poisson averaged models (both weighted and non-weighted) were also updated every 24 hours. A sliding window of 4 hours (H=8) was considered in the ensemble.

Table 1: Taxi Services Volume (Per Daytype/Shift)

Daytype Group	Total Services Emerged	Averaged Service Demand per Shift		
		0am to 8am	8am to 4pm	4pm to 0am
Workdays	957265	935	2055	1422
Weekends	226504	947	2411	1909
All Daytypes	1380153	1029	2023	1503

Table 2: Taxi Services Volume(Per Driver/Cruise Time)

	Services per Driver	Total Cruise Time (minutes)
Maximum	6751	71750
Minimum	100	643
Mean	2679	33132
Std. Dev.	1162	13902

5.2 Traffic Simulator: An Online Test-Bed

The DIVERT [Conceicao *et al.*, 2008] is a high-performance traffic simulator framework which uses a realistic microscopic mobility model. The main advantage of this framework when facing others is the easiness to create new simulation modules efficiently. Hence, we have created a new model that simulates the real behavior of a taxi fleet. Upon a request, a central entity elects one taxi to do the requested service. Once the service is finished, the same entity recommends a new taxi-stand for the taxi to go to and wait for a new service.

This framework was employed as an online test-bed for our Recommendation System. Firstly, the realistic map of the city of Porto - containing the real road network topology and the exact location of the 63 taxi stands in the city – was loaded.). Secondly, we fed the framework with a service log (i.e. a time-dependent origin-destination matrix) correspondent to the studied period. However, we just accessed the log of one out of the two running fleets in Porto (the largest one, with 441 vehicles). To simulate a scenario similar to our own, we divided this fleet into two using a ratio close to real one (60% for the *fleet* A1 and 40% to the *fleet* B1). The services dispatched from the central were also divided in the same proportion while the services demanded in each taxi stand will be the same. The *fleet* B1 will use the most common and traditional way to choose the best taxi-stand: it will go to the nearest taxi stand of each drop-off location (i.e. after a drop-off, each driver has to head to a specific taxi stand of its own choice). However, the *fleet* A1 will use our Recommendation System to do an *informed driving*, which considers multiple

variables – like the number of taxis in each stand or the demand prediction on them - to support this important decision. Finally, we ran the simulation and we extract the metrics for each fleet. The framework is used to calculate the optimal paths between the taxi stand and the passenger location and the dependent behavior of the fleets (the location of each vehicle will affect the way they get the services). Our main goal is to simulate a real scenario behavior and its competitive characteristics while we are testing the Recommendation System. It is important to notice that both fleets would get similar results if they did not use any Recommendation System. We also highlight that the vehicles will remain parked in the stand waiting for a service whenever the time it takes to appear. In this case, we consider the maximum threshold of 120 minutes that is deeply detailed in the following section, along with the remaining evaluation metrics.

5.3 Evaluation Methods

We used the data obtained from the last four months to evaluate our both experimental setups (where 506873 services emerged). Firstly, we present two error measurements which were employed to evaluate our output: one from the literature and another from our own specifically adapted to our current problem. Secondly, we detail the two performance metrics used to evaluate our recommendation models.

Consider $R_k = \{R_{k,0}, R_{k,1}, ..., R_{k,t}\}$ to be a discrete time series (aggregation period of P-minutes) with the number of services predicted for a taxi stand of interest k in the period $\{1, t\}$ and $X_k = \{X_{k,0}, X_{k,1}, ..., X_{k,t}\}$ the number of services actually emerged in the same conditions. The (1) Symmetric Mean Percentage Error (*sMAPE*) is a well-known metric to evaluate the success of time series forecast models. However, this metric can be too intolerant with small magnitude errors (e.g. if two services are predicted on a given period for a taxi stand of interest but no one actually emerges, the error within that period would be 1). Then, we propose to also use an adapted version of Normalized Mean Absolute Error (NMAE).

The (2) Average Weighted Error (*AVE*) is a metric of our own based on the NMAE. We defined it as

$$AVE' = \sum_{i=1}^{t} \frac{\theta_{k,i} * X_{k,i}}{\sigma_{k,i} * \psi_k} \tag{6}$$

$$\sigma_{k,i} = \left\{ \begin{array}{ll} X_{k,i} & \text{if } X_{k,i} > 0 \\ 1 & \text{if } X_{k,i} = 0 \end{array} \right\} \tag{7}$$

$$\theta_{k,i} = \left\{ \begin{array}{ll} |R_{k,i} - X_{k,i}| & \text{if } X_{k,i} > \text{th} \\ 0 & \text{if } X_{k,i} \leq \text{th} \end{array} \right\} \tag{8}$$

$$\psi_k = \sum_{i=1}^{t} X_{k,i}, AVE = \left\{ \begin{array}{ll} AVE' & \text{if AVE'} \leq 1 \\ 1 & \text{if AVE'} > 1 \end{array} \right\} \tag{9}$$

where ψ_k is the total of services emerged at the taxi stand k during the time period $\{1, t\}$. The main feature about this metric is to weight the error in each period by the number of real events actually emerged (i.e. the errors on periods where more services were actually demanded are more relevant than the remaining ones).

Both metrics are focused just on one time series for a given taxi stand. However, the results presented below use an averaged error measured based on all stands series – *GA*. Consider β to be an error metric of interest. AG_β is an aggregated metric given by a weighted average of the error in all stands. It is formally presented in the following equation.

$$AG_\beta = \sum_{k=1}^{N} \frac{GA_{\beta,k} * \psi_k}{\mu}, \mu = \sum_{k=1}^{N} \psi_k \tag{10}$$

We considered three performance metrics in the evaluation of our recommendation models: (1) the *Waiting Time* (WT) and (2) the *Vacant Running Distance* (VRD) and the number of *No Services* (NS). The Waiting Time is the total time that a driver takes between a drop-off and a pick-up (i.e. to leave a stand with a passenger or to get one in his/her current location). The *Vacant Running Distance* is the distance that a driver does to get into a stand after a drop-off (i.e.: without any passenger inside). Independently on the time measured on the simulation, we always consider a maximum threshold of 120 minutes to the *Waiting Time*. The No Service metric is a ratio between the number of times that a taxi parked on a stand had a waiting time greater than the 120 minutes threshold and the number of services effectively dispatched by the respective fleet.

5.4 Results

Firstly, we present the results obtained by the online experiments done with the predictive models. The error measured for each model is highlighted in Table 3 and Table 4. The results are firstly presented per shift and then globally. These error values were aggregated using the AG_β previously defined.

Secondly, the values calculated for our performance metrics using the traffic simulator previously described are detailed in the Table 5. The *fleet A1* used the Recommendation Model 1 (RS1) while the *B1* uses the common expected behavior (previously defined). Distinct metrics values are presented for the two using different aggregations like the arithmetic mean (i.e. average), the median and the standard deviation. The *No Services* ratio is also displayed.

6 Final Remarks

In this paper, we present a **novel application of time series forecasting techniques** to improve the taxi driver **mobility intelligence**. We did it in three distinct steps: firstly (1) we mined both GPS and event signals emitted by a company operating in Porto, Portugal (where the passenger demand is

Table 3: Error Measured on the Models using AVE

Model	**Periods**			
	00h−08h	**08h−16h**	**16h−00h**	**24h**
Poisson Mean	21.28%	24.88%	22.88%	23.43%
W. Poisson Mean	23.32%	28.37%	26.77%	26.74%
ARIMA	20.85%	26.12%	22.92%	20.91%
Ensemble	**14.37%**	**18.18%**	**17.19%**	**15.89%**

Table 4: Error Measured on the Models using $sMAPE$

Model	Periods			
	00h−08h	08h−16h	16h−00h	24h
Poisson Mean	15.09%	19.20%	17.51%	16.84%
W. Poisson Mean	17.32%	20.66%	19.88%	18.47%
ARIMA	16.81%	18.59%	17.85%	18.51%
Ensemble	**14.37%**	**18.18%**	**17.19%**	**15.89%**

Table 5: An Analysis on the Recommendation Performance

Performance Metrics	A1(RS)	B1(common)
Average WT	38.98	40.84
Median WT	26.29	27.92
Std. Dev. WT	33.79	33.22
Average VRD	3.27	1.06
Median VRD	2.80	0.98
Std. Dev. VRD	2.53	0.54
No Service (%)	**11.08%**	**19.26%**

lower than the vacant taxis). Secondly, we predicted - **in a real-time experiment** - the distribution of the taxi-passenger demand for the 63 taxi stands at 30-minute period intervals. Finally, we recreated the scenario running in Porto, where two fleets (the fleet A and B, which contain 441 and 250 vehicles, respectively) compete to get as many services as possible. We did it using a **traffic simulation framework** fed by the real services historical log of the largest operating fleet. One of the fleets used our Recommendation System for the Taxi Stand choice problem while the other one just picked the stand using a baseline model corresponding to the driver common behavior in similar situations.

Our predictive model demonstrated a more than satisfactory performance, anticipating in real time the spatial distribution of the passenger demand with an error of just 20%. We believe that **this model is a true novelty and a major contribution** to the area through its online adapting characteristics:

- It takes advantage of the ubiquitous characteristics of a taxi communicational network, assembling the experience and the knowledge of all vehicles/drivers while they usually use just their own;
- It simultaneously uses long-term, mid-term and short term historical data as a learning base;
- It rapidly produces real-time short-term predictions of the demand, which can truly **improve drivers' mobility intelligence** and consequently, their profit.

This approach meets no parallel in the literature also by its test-bed: the models were tested in a streaming environment, while the state-of-art presents mainly offline experimental setups. Our simulation results demonstrated that such **informed driving** can truly improve the drivers' mobility intelligence: the *fleet A1* had an *Average Waiting Time* 5% lower than its competitor – even if it has a larger fleet. We also highlight the reduction of the *No Service* ratio in 50% while the *Vacant Running Time* faced an increase. It is important to state that this Recommendation System is focused on a Scenario like our own – two or more competitors operating in a medium/large city where the demand is lower than the number of running vehicles. Its main goal is to recommend a stand where a service will rapidly emerge – even if this stand is far away. The idea is to be in a position able to pick-up the emerging service demand before the remaining competition. This factor can provoke a slight increase on the Vacant Running Time but it will also reduce the usually large Waiting Times to pick-up passengers. Other scenarios may require a distinct calibration of the model to account different needs/goals.

Acknowledgments

The authors would like to thank to Geolink and to its team for the data supplied to this work. This work was supported by the projects DRIVE-IN: "Distributed Routing and Infotainment through Vehicular Internet-working", VTL: "Virtual Traffic Lights" and KDUS: "Knowledge Discovery from Ubiquitous Data Streams" under the Grants CMU-PT/NGN/0052/2008, PTDC/EIA-CCO/118114/2010, PTDC/EIA-EIA/098355/2008, respectively, and also by ERDF - European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness), by the Portuguese Funds through the FCT(Portuguese Foundation for Science and Technology) within project FCOMP-01-0124-FEDER-022701.

References

[Box *et al.*, 1976] G. Box, G. Jenkins, and G. Reinsel. *Time series analysis*. Holden-day San Francisco, 1976.

[Chang *et al.*, 2010] H. Chang, Y. Tai, and J. Hsu. Context-aware taxi demand hotspots prediction. *International Journal of Business Intelligence and Data Mining*, 5(1):3–18, 2010.

[Conceicao *et al.*, 2008] Hugo Conceicao, Luis Damas, Michel Ferreira, and Joao Barros. Large-scale simulation of v2v environments. In *Proceedings of the 2008 ACM symposium on Applied computing*, pages 28–33. ACM, 2008.

[Contreras *et al.*, 2003] J. Contreras, R. Espinola, F. J. Nogales, and A. J. Conejo. Arima models to predict next-day electricity prices. *IEEE Transactions on Power Systems*, 18(3):1014–1020, 2003.

[Cryer and Chan, 2008] J. Cryer and K. Chan. *Time Series Analysis with Applications in R*. Springer, USA, 2008.

[Deng and Ji, 2011] Z. Deng and M. Ji. Spatiotemporal structure of taxi services in shanghai: Using exploratory spatial data analysis. In *Geoinformatics, 2011 19th International Conference on*, pages 1–5. IEEE, 2011.

[Holt, 2004] Charles Holt. Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting*, 20(1):5–10, 2004.

[Kaltenbrunner *et al.*, 2010] Andreas Kaltenbrunner, Rodrigo Meza, Jens Grivolla, Joan Codina, and Rafael

Banchs. Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system. *Pervasive and Mobile Computing*, 6(4):455–466, 2010.

[Lee *et al.*, 2008] J. Lee, I. Shin, and G.L. Park. Analysis of the passenger pick-up pattern for taxi location recommendation. In *Fourth International Conference on Networked Computing and Advanced Information Management (NCM'08)*, volume 1, pages 199–204. IEEE, 2008.

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

[Li *et al.*, 2012] Xiaolong Li, Gang Pan, Zhaohui Wu, Guande Qi, Shijian Li, Daqing Zhang, Wangsheng Zhang, and Zonghui Wang. Prediction of urban human mobility using large-scale taxi traces and its applications. *Frontiers of Computer Science in China*, 6(1):111–121, 2012.

[Liu *et al.*, 2009] L. Liu, C. Andris, A. Biderman, and C. Ratti. Uncovering taxi drivers mobility intelligence through his trace. *IEEE Pervasive Computing*, 160:1–17, 2009.

[Min and Wynter, 2011] W. Min and L. Wynter. Real-time road traffic prediction with spatio-temporal correlations. *Transportation Research Part C: Emerging Technologies*, 19(4):606–616, 2011.

[Moreira-Matias *et al.*, 2012] Luis Moreira-Matias, Joao Gama, Michel Ferreira, Joao Mendes-Moreira, and Luis Damas. Online predictive model for taxi services. In *Advances in Intelligent Data Analysis XI*, volume 7619 of *LNCS*, pages 230–240. Springer Berlin Heidelberg, 2012.

[Phithakkitnukoon *et al.*, 2010] S. Phithakkitnukoon, M. Veloso, C. Bento, A. Biderman, and C. Ratti. Taxi-aware map: identifying and predicting vacant taxis in the city. *Ambient Intelligence*, 6439:86–95, 2010.

[Wang *et al.*, 2003] H. Wang, W. Fan, P.S. Yu, and J. Han. Mining concept-drifting data streams using ensemble classifiers. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 226–235. ACM, 2003.

[Yeasmin and Rob, 1999] Khandakar Yeasmin and J. Hyndman Rob. *Automatic Time Series Forecasting: The forecast Package for R*, 1999.

[Yuan *et al.*, 2010] J. Yuan, Y. Zheng, C. Zhang, W. Xie, X. Xie, G. Sun, and Y. Huang. T-drive: driving directions based on taxi trajectories. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 99–108. ACM, 2010.

[Yue *et al.*, 2009] Y. Yue, Y. Zhuang, Q. Li, and Q. Mao. Mining time-dependent attractive areas and movement patterns from taxi trajectory data. In *Geoinformatics, 2009 17th International Conference on*, pages 1–6. IEEE, 2009.

[Zhang, 2003] G.Peter Zhang. Time series forecasting using a hybrid arima and neural network model. *Neurocomputing*, 50:159–175, 2003.