2012 15th International IEEE Conference on Intelligent Transportation Systems
Anchorage, Alaska, USA, September 16-19, 2012

A Predictive Model for the Passenger Demand on a Taxi Network

Luis Moreira-Matias, João Gama, Michel Ferreira, Luís Damas

Abstract— **In the last decade, the real-time vehicle location systems attracted everyone attention for the new kind of rich spatio-temporal information. The fast processing of this large amount of information is a growing and explosive challenge. Taxi companies are already exploring such information in efficient taxi dispatching and time-saving route finding. In this paper, we propose a novel methodology to produce online short term predictions on the passenger demand spatial distribution over 63 taxi stands in the city of Porto, Portugal. We did so using time series forecasting techniques to the processed events constantly communicated for 441 taxi vehicles. Our tests - using 4 months of real data - demonstrated that this model is a true major contribution to the driver mobility intelligence: 76% of the 86411 demanded taxi services were accurately forecasted in a 30 minutes time horizon.**

I. Introduction

Taxis are an important mean of transportation which offers a comfortable and direct service as complement of mass transportation services or as an individual speedy transportation demand (usual or not) [1]. In the last decade, the real-time vehicle location systems (using GPS - Global Positioning System and wireless communication features) attracted the attention of both taxi companies and researchers for the new kind of rich spatio-temporal information. Online systems for efficient taxi dispatching and time-saving route finding (among others) were developed to improve taxi service reliability.

One of the main obstacles to the passenger picking are the existing regulations limiting their activity, namely, the licensed areas and applicable fares. Thus, the equilibrium between the passenger demand and taxi drivers search for customers is fundamental to maximize profit. H. Yang and S. C. Wong presented relevant mathematical models to express this equilibrium in distinct contexts [1, 2]. An equilibrium fault may lead to one of two scenarios: (Scenario 1) excess of vacant vehicles and excessive competition or (Scenario 2) larger passenger waiting times and lower taxi reliability. Our focus is just on first scenario, even if the second is the most studied one [3-5].

The taxi driver mobility intelligence is a crucial feature to maximize both profit and reliability on both scenarios. There are few works focused on this topic using GPS real data. The work presented in [5] used spatial-based clustering applied to GPS historical data from a taxi network (Scenario 2) running in a large urban zone. Their goal was to discover the passenger demand patterns over time. A similar work was introduced on [6] where a 3D clustering technique is used to analyze the spatiotemporal patterns for both top and ordinary drivers. This work main focus was to reveal top driver mobility intelligence. Recently, an innovative study was presented in [3] to validate the triplet Time-Location-Strategy as the key features to build a good passenger finding strategy. They used a L1-Norm-SVM as a feature selection tool to discover both efficient and inefficient passenger finding strategies based on a large-scale GPS dataset from a taxi network running on a large city in China. They also made an empirical study on the impact of the selected features and its conclusions were validated by the feature selection tool.

All reported works have four common characteristics: (1) they used GPS real data to study which factors affect meaningfully the taxi driver mobility intelligence and consequently its choice about the best route to take to maximize the probabilities to find suitable passengers; (2) they were tested in a Scenario 2 city, where the passenger demand is usually larger than the number of taxis available; (3) they assumed that the drivers can choose their routes freely and without any regulation and (4) their conclusions were obtained offline. Despite the useful insights discovered, the reported frameworks just provide offline information. In other words, those are not ubiquitous frameworks to aid in the taxi driver decision making…even if the dataset used was a stream one. Moreover, there are several urban areas and countries where the taxi companies or the industry regulations do not allow the vacant taxis to cruise the roads *randomly* to get passengers: after the passenger drop, the drivers are forced to pick a route to one of the local taxi stands to wait to pick up another passenger (i.e. a service – this terminology is used from now on). Even if the absence of regulation is in place, the rising cost of fuel is clearly disallowing the economic viability of cruising strategies to get passengers.

In our work, we focused on the choice problem about which is the best taxi stand to go after a passenger drop-off

Manuscript submitted March 15, 2012. This work was supported by the

Luis Moreira-Matias is with the DEI, FEUP, Universidade do Porto, Rua Dr. Roberto Frias, s/n 4200-465 Porto - Portugal and with LIAAD-INESC Porto L.A. Rua de Ceuta, 118, 6º; 4050-190 Porto - Portugal (corresponding author to provide phone: 00351-91-4221647; e-mail: luis.matias@fe.up.pt).

João Gama is with the FEP, Universidade do Porto, Rua Dr. Roberto Frias, s/n 4200-465 Porto - Portugal and with LIAAD-INESC Porto L.A. Rua de Ceuta, 118, 6º; 4050-190 Porto - Portugal (e-mail: jgama@fep.up.pt).

Michel Ferreira is with the Instituto de Telecomunicações, Departamento de Ciência de Computadores, Faculdade de Ciências, Universidade do Porto, 4169-007 Porto, Portugal (e-mail: michel@dcc.fc.up.pt).

Luis Damas is with Geolink, Pct. Adelino Amaro Costa, 772 4º Dto, 4050-012 Porto – Portugal (e-mail: luis@geolink.pt)

978-1-4673-3063-3/12/$31.00 ©2012 IEEE

in a given location and time. This choice is related with four key factors: the expected price for a service over time, the distance/cost relation with each stand, how many taxis are already waiting in each stand and the passenger demand for each stand over time. In this paper, we specifically addressed the issues related with this last factor: we present an ubiquitous model to predict the number of services on a taxi network over space (taxi stand) for a short-time horizon of *P*-minutes. Based on historical GPS location and service data (passenger drop-off and pick-up), time series histograms are built for each stand containing the number of services with an aggregation of *P*-minutes. The predictive model was developed adapting well-known time series forecasting techniques such as time varying Poisson model [7] and ARIMA (Autoregressive Integrated Moving Average) [8] to our problem. Our goal is to predict at the instant *t* how many services will be demanded during the period [*t, t+P*] at each existent taxi stand, reusing the real service count on [*t, t+P*] extracted from the data to do the same for the instant *t+P* and so on (i.e. the framework run continuously in a stream). To the best of our knowledge, such approach has no parallel in the literature.

We applied our model to data from a large-sized taxi network containing a total of 63 taxi stands and 441 vehicles running on the city of Porto, Portugal (Scenario 1). In this city, the taxi drivers must pick a route to one of the existing stands after a passenger drop-off. Even so, the vacant ones can pick-up passengers on the street while cruising to a taxi stand. However, the parked vehicles have a higher priority in the dispatch system than the cruising ones. Our study just uses as input/output the services got directly on the stands or automatically dispatched to the parked vehicles, ignoring the remaining ones. We did so because the passenger demand in each taxi stand over is the main feature to aid the taxi drivers decision, since it represents 76% of the total number of services.

Our test-bed was a computational stream simulation running offline. The first 13 weeks of the dataset were used as training set and the last 3 were used as input for our stream-type test-bed (i.e. simulating the service demands that would arrive continuously in a stream). The results obtained were promising: our model accurately predicted more than 76% of the services that actually emerged.

The remainder of the paper is structured as follows. Section 2 formally describes our model. Section 3 describes how we acquired and preprocessed the dataset used as well as some statistics about it. Section 4 describes how we tested the methodology in a concrete scenario. Firstly, we introduce the evaluation metrics of our model and the experimental setup used. Then, we present the obtained results. Section 5 concludes and describes the future work we intend to carry on.

II. THE MODEL

Let $\boldsymbol{S} = \{\boldsymbol{s_1}, \boldsymbol{s_2}, \dots, \boldsymbol{s_N}\}$ be the set of $\boldsymbol{N}$ taxi stands of interest and $\boldsymbol{D} = \{\boldsymbol{d_1}, \boldsymbol{d_2}, \dots, \boldsymbol{d_j}\}$ be a set of $\boldsymbol{j}$ possible passenger destinations. Our problem is to choose the best taxi stand at instant $\boldsymbol{t}$ according with our forecast about passenger demand distribution over the time stands for the period [$\boldsymbol{t}$, $\boldsymbol{t+P}$], as is illustrated in Fig. 1.

Let $\boldsymbol{X_k} = \{\boldsymbol{X_{k,0}}, \boldsymbol{X_{k,1}}, \dots, \boldsymbol{X_{k,t}}\}$ be a time series for the number of demanded services at a taxi stand $\boldsymbol{k}$. Our goal is to build a model to determine the set of service count $\boldsymbol{X_{k,t+1}}$ forthe instant $\boldsymbol{t+1}$ and for all taxi stands $\boldsymbol{k \, \epsilon \, \{1, N\}}$. To do so, we propose three distinct short-term prediction models and a well-known data stream ensemble framework to use them all. We formally describe those models along this section.

A. *Time Varying Poisson Model*

The following section presents a model firstly proposed in [7]. The demand on taxi services exhibit, like other transportation means [9], a periodicity in time on a daily basis that reflects the patterns of the underlying human activity, making the data appear non-homogeneous [7]. Fig. 2 illustrates a one month taxi service analysis extracted from our dataset that illustrates this periodicity (the dataset is described in detail in the Section III).

Let the probability to have *n* taxi assigns in a determined time period – ***P(n)*** - follow a **Poisson distribution**. We can define it using the following equation

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

where λ represents the rate (averaged number of the demand on taxi services) in a fixed time interval. However, in this specific problem, the rate λ is not constant but time-variant. So, we adapt it as a function of time, i.e. $\lambda(t)$, transforming the Poisson distribution into a nonhomogeneous one. Let $\lambda(t)$ be defined as follow

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

where $d(t)$ represents the weekday {1=Sunday,2=Monday,…}; $h(t)$ the period in which time *t* falls (e.g. the time 00:31 is contained in the period 2 if we consider 30-minute periods).

It requires the validity of both equations

$$\sum_{\mathrm{i}=1}^{7} \delta_{\mathrm{i}} = 7, \qquad (3)$$

$$\sum_{\mathrm{i}=1}^{\mathrm{D}} \eta_{\mathrm{d,i}} = \mathrm{D} \quad \forall \mathrm{d}, \qquad (4)$$

where D is the number of time intervals in a day. To ease the interpretation of these equations, we can define the remaining symbols as follows:

- λ_0 is the average (i.e. expected) rate of the Poisson process over a full week;
- δ_{i} is the relative change for the day $\boldsymbol{i}$ (Saturday have lower day rates than Tuesdays);
- $\eta_{\mathrm{j,i}}$ is the relative change for the period $\boldsymbol{i}$ on the day $\boldsymbol{j}$ (the peak hours);

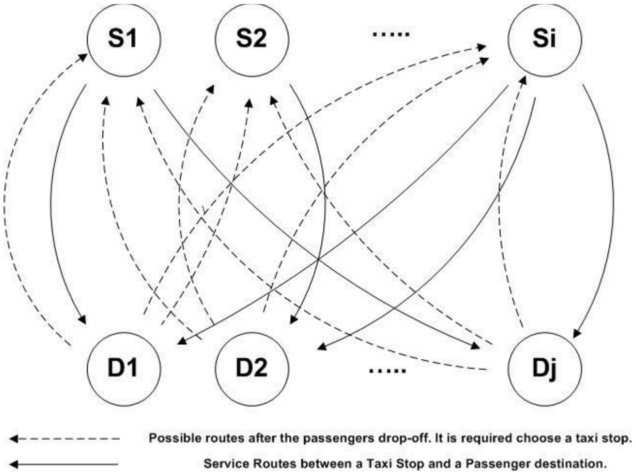
S1
S2
.....
Si
D1
D2
.....
Dj
Possible routes after the passengers drop-off. It is required choose a taxi stop.
Service Routes between a Taxi Stop and a Passenger destination.

Figure 1. A schema to illustrate our problem.

- $\lambda(t)$ is a discrete function representing the expected demand on taxi services distribution over time for a taxi stand of interest ***k***.

B. Weighted Time Varying Poisson Model

The model previously presented can be faced as a time-dependent average. However, it is not guaranteed that every taxi stand have highly regular passenger demand: actually, the demand in many stands can be often corrupted by seasonal bursty periods of expected events like large crowd events, weather changes and so on (an illustrated example is presented in Fig. 3).

To face this specific seasonal issue, we propose an weighted average model based in the one already presented before: our goal is to increase the relevance of the demand pattern observed in the last week comparing to the patterns observed several weeks ago (e.g. what happened in the last 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 this kind of problems: the Exponential Smoothing [10]. We can define ω as following

$$\omega = \alpha * \{1, (1-\alpha), (1-\alpha)^2, \ldots, (1-\alpha)^{\gamma-1}\}, \qquad (5)$$

where γ is the number of historical periods considered in the initial average, α is the smoothing factor (i.e. a user-defined parameter) and $0 < \alpha < 1$.

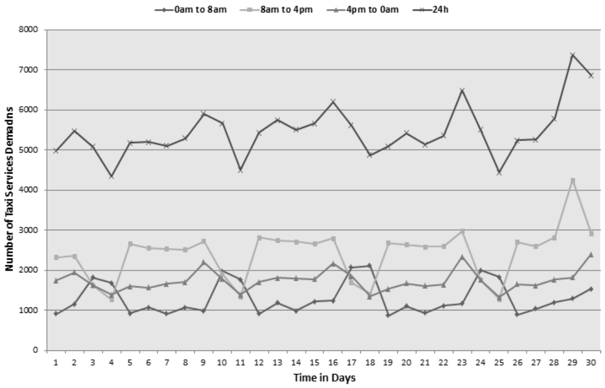

Figure 2. One month data analysis (total and per driver shift).

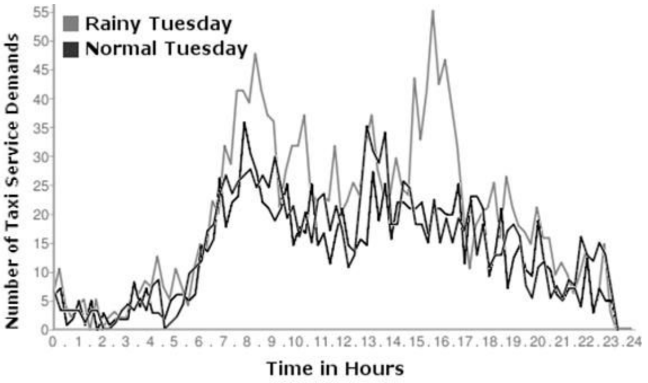

Figure 3. Taxi Services daily profiles aggregated in 15-minutes periods. Two correspond to typical Tuesdays and the other to a rainy one.

C. Autoregressive Integrated Moving Average Model

The two last models assume the existence of a regular (seasonal or not) periodicity in the taxi service passenger demand (i.e. the demand in one taxi stand in a regular Tuesday during a certain period will be highly similar to the demand verified during the same period in other Tuesdays).

However, the demand can be different than this from stand to stand and the existence of other periodicities (e.g.Thursdays with the Tuesdays, mornings with the evenings and so on) should be explored to achieve a better accuracy prediction.

The Autoregressive Integrated Moving Average Model (ARIMA) [8] is a well-known methodology to both model and forecast univariate time series data such as traffic flow data [11] and other short term prediction problems like our own. The ARIMA main advantage facing other algorithms is its versatility to represent very different types of time series: the autoregressive (AR) ones, the moving average ones (MA) and a combination of those two (ARMA). A brief presentation of one of the simplest ARIMA models (for non-seasonal stationary time series) is enunciated below following the existing description in [12] (however, our framework can also detect both seasonal and non-stationary ones). For a more detailed discussion, the reader should consult a comprehensive time series forecasting text such as Chapters 4 and 5 in [13].

In an autoregressive integrated moving average model, the future value of a variable is assumed to be a linear function of several past observations and random errors. We can formulate the underlying process that generate the time series (taxi service over time for a given stand ***k***) as

$$\begin{aligned} \mathrm{R}_{\mathrm{k,t}} = {} & \theta_0 + \phi_1 \mathrm{X}_{\mathrm{k,t-1}} + \phi_2 \mathrm{X}_{\mathrm{k,t-2}} + \cdots + \phi_p \mathrm{X}_{\mathrm{k,t-p}} \\ & + \varepsilon_{\mathrm{k,t}} - \theta_1 \mathrm{X}_{\mathrm{k,t-1}} - \theta_2 \mathrm{X}_{\mathrm{k,t-2}} - \cdots - \theta_q \mathrm{X}_{\mathrm{k,t-q}} \qquad (6) \end{aligned}$$

where $\mathrm{R}_{\mathrm{k,t}}$ and $\varepsilon_{k,t}$ are the actual value and the random error at time period ***t***, respectively; $\phi_l (l = 1,2, \ldots, p)$ and $\theta_m (m = 0,1,2, \ldots, q)$ are the model parameters/weights while p and q are positive integers often referred as the order of the model. Both order and weights can be inferred from the historical time series using both the autocorrelation and partial autocorrelation functions like it was introduced by

Box and Jenkins in [14]. They are useful to detect if the signal is periodic and, most important, which are the frequencies of these periodicities. A study on time series from the demand on taxi services in one of the busiest taxi stands is displayed on Fig. 4.

D. *Sliding Window Ensemble Framework*

In the last decade, regression and classification tasks on streams attracted the community attention due to its ubiquitous characteristics. The ensembles of such models were specifically focused due to the challenge related with. One of the most popular models is the weighted ensemble [15]. The model we propose below is based on this one.

Let $M = \{M_1, M_2, \dots, M_z\}$ be a set of z models of interest to model a given time series and $M_t = \{M_{1t}, M_{2t}, \dots, M_{zt}\}$ be the set of forecasted values to the next period on the interval $\boldsymbol{t}$ by those models. The ensemble forecast E_t is obtained as

$$\mathrm{E_t} = \sum_{i=1}^{z} \frac{\mathrm{M_{it}}}{\beta}, \beta = \sum_{i=1}^{z} \rho_{iH} \qquad (7)$$

where ρ_{iH} is the forecasting accuracy obtained for the model $\mathrm{M_i}$ in the periods contained on the time window/period $[\mathrm{t} - \mathrm{H}, \mathrm{t}]$ (H is a user-defined parameter to define the window size). As the information is arriving in a continuous manner for the next periods $\{t, t+1, t+2, \dots.\}$ the window will also **slide** to determine how are the models performing in the **last H periods**. To calculate such accuracy, we used an well-known time series forecasting error metric: the Symmetric Mean Percentage Error (*sMAPE*) [16].

III. Data Acquisition and Preprocessing

In this work, we studied the taxi driver mobility intelligence of one company operating in the city of Porto, Portugal. This city is the center of a medium size urban area (1.3 million of habitants) where the passenger demand is inferior to the number of running vacant taxis, provoking a huge competition within both companies and drivers - according to a recent aerial survey of the road traffic of the city [17], taxis represent 4% of the running vehicles during a non-rush hour period. The existing regulations force the drivers not to run *randomly* searching for passenger but to choose a specific taxi stand out of the 63 existing in the city – see Fig. 5 to observe its spatial distribution - to go park and wait for the next service immediately after the last passenger drop-off. There are three main ways to pick-up a passenger in: (I) a passenger goes to a taxi stand and pick-up a taxi – the regulations also force the passengers to pick-up the first taxi in the line (First In, First Out); (II) a passenger calls to the taxi network central and demand a taxi for a specific location/time – the parked taxis have priority over the running vacant ones in the central taxi dispatch system; (III) a passenger pick a vacant taxi while it is going to a taxi stand, in any street.

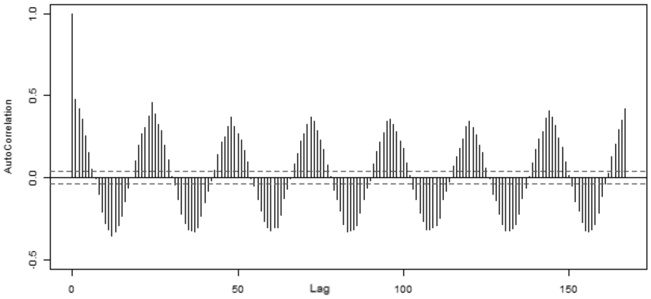

Figure 4. Autocorrelation profile for data about the demand on taxi service (13 weeks) obtained from one of the busiest taxi stands in the city (periods of 60-minutes). The x-axis has the different period lags studied and the y-axis has the correlation within the signal. Note the peaks for each 12h periods.

In this section, we describe the studied company and the data acquisition process and the preprocessing applied to it.

A. *Data Acquisition*

The data was continuously acquired using the telematics installed in each one of the 441 running vehicles (GPS and wireless communication) in the studied company. These vehicles usually run in one out of three 8h shifts: midnight to 8am, 8am-4pm and 4pm to midnight. Each data chunk arrives with the following six attributes: (1) TYPE – it is relative to the type of event reported and it has four possible values: *busy* – the driver picked-up a passenger; *assign* – the dispatch central assigned a service previously demanded; *free* – the driver dropped-off a passenger and *park* – the driver parked in a taxi stand. The attribute (2) STOP is an integer with the ID of the related taxi stand. The attribute (3) TIMESTAMP is the date/time in seconds of the event and the attribute (4) TAXI is the driver code; the attributes (5) and (6) are the LATITUDE and the LONGITUDE corresponding to the acquired GPS position.

This data was acquired for non-stop period of 16 weeks. Our study just uses as input/output the services got directly on the stands or automatically dispatched to the parked vehicles (more details in the section below). We did so because the passenger demand in each taxi stand over is the main feature to aid the taxi drivers' decision.

B. *Preprocessing and Data Analysis*

As preprocessing, a time series of taxi demand services aggregated for period of P-minutes was developed. There are three types of accounted events: (1) the *busy* got directly in a taxi stand; (2) the *assign* got directly to a taxi parked in a taxi stand and (3) the *busy* got while a vacant taxi is cruising. The type 1 events are accounted directly as the type 2. However, for each type 2 event, the system receives a *busy* event a few minutes later – as soon as the driver

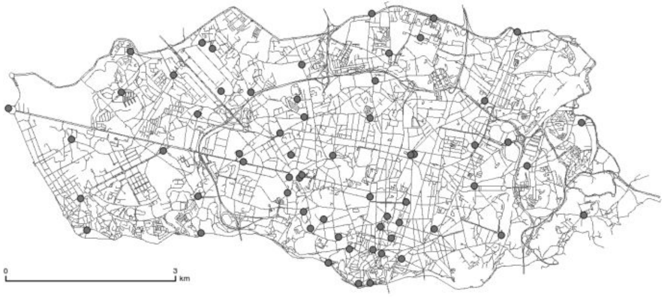

Figure 5. Taxi Stand spatial distribution over the city of Porto, Portugal.

effectively picked-up the passenger – that is ignored by our system. The events of type 3 are ignored unless they occur in a radius of *W* meters of a taxi stand (where *W* is a user defined parameter). If it does, it is considered as a type 1 event related with the nearest taxi stand according the defined criteria. We did so because many regulations disallows to pick-up passengers in a pre-defined radius of a stop (in Porto is defined a 50m radius).

Table I details the number of taxi services demanded from any taxi stand per daily shift and day type. Table II has information about all services (independently of the pick-up place) per taxi and duration. The *service* column in Table II represents the number of services picked-up by the taxi drivers, while the second one is related with the time distance of each service done.

Additionally, we could state that the central service assignment is 24% of the total service (*versus* the 76% of the one demanded directly in the street) while 77% of the service is demanded directly to taxis parked in a taxi stand (and 23% is picked-up while they are cruising). The average waiting time (to pick-up passengers) of a taxi parked in a taxi stand is 42 minutes while the average time distance for a service is only 11 minutes and 12 seconds.

The data in the tables highlight that, despite the regularity exhibited in the service (especially on the weekends), there are huge mobility intelligence discrepancies between the drivers (i.e. a large variance in both number and time distance of the services).

IV. EXPERIMENTAL RESULTS

In this section, we firstly describe the experimental setup developed to test our model on the available data. Secondly, we present and discuss the results achieved.

A. Experimental Setup

Our model produces an online forecast for the demand on taxi services in all taxi stands at each P-minutes period. However, we just used an offline continuous simulation to test it. The scripts used were developed using the R statistical software [18]. The pre-defined functions used and the values set for the models parameters are detailed along this section.

It was set an aggregation period of 30 minutes (i.e. a new forecast is produced each 30 minutes; *P=30*) and a radius of100 meters (*W* = 100 > 50 defined by the existing regulations). This aggregation was set according the average waiting time in taxi stand (< 42 minutes).

The previously described dataset was divided into a training/test set for periods of 13/3 weeks, respectively. Both training and test set were composed by one time series per taxi stand and each value tested on the period *t* was merged to the training set to generate the forecast on the period *t+1* (i.e. the real number of taxi services count for each taxi stand along 30 minutes are considered for the next period forecast and so on).

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 [19] - *auto-arima* – with the default parameters. The weights/parameters are specifically fit for each period using the function *arima* from the built-in R package [*stats*].

TABLE I. TAXI SERVICES VOLUME (PER DAYTYPE/SHIFT)

	Total Services Occurred	**Averaged Service Demand per Shift**		
		0am to 8am	**8am to 4pm**	**4pm to 0am**
Workdays	309369	900	2344	2383
Weekends	92297	1483	1354	1428
Total	**401666**	**2383**	**3698**	**3811**

TABLE II. TAXI SERVICES VOLUME (PER TAXI/DURATION)

	Services	**Time Running Busy** (minutes)
Máx.	4020	32987
Min.	38	384
Mean Total	1104	13761
Std. Dev. Total	425	5004
Mean t<30m	1058	
Std. Dev. t<30m	413	

The time-varying Poisson averaged models (both weighted and non-weighted) were updated each 24 hours. A sliding window of 4 hours (H=8) was considered in the ensemble. The accuracy of each model was measured using the metric also proposed to weight each model in the ensemble – *sMAPE*. Distinct results for two distinct values (0.4 and 0.5) of the parameter *alpha* (α in the weighted average) are presented below.

B. Results

A total of 86411 services were tested. The accuracy measured for each one the models are presented in Table III and Table IV. The results are firstly presented per shift and then globally. The values presented below are calculated through an average weight of the accuracy obtained in each one of the time series (i.e. the accuracy of the forecast on the demand for taxi services on each one of the 63 taxi stands). Each accuracy was weighted according to the number of services demanded on the correspondent taxi stand along all the test period.

Each model presents accuracy above the 74% in both tables. The W. Poisson Mean and the Ensemble are the only affected by the changes on the *alpha* (α) parameter. The sliding window ensemble is always the best model in every shift and period considered, with an accuracy superior to 76% (65672 of the 86411 total taxi services were correctly forecasted in both time and space using an aggregation of 30-minutes).

V. CONCLUSIONS AND FUTURE WORK

In this paper, we presented a **novel application of time**

series forecasting techniques to **improve the taxi driver mobility intelligence**. We did so by transforming both GPS and event signals emitted by 441 taxis of a company lower operating in Porto, Portugal (where the passenger demand is than the number of vacant taxis) into time series of interest to use firstly (1) as an offline learning base to our model and secondly (2) as an online test framework. As result, our model was able to predict the passenger demand on taxi services on each one of the 63 taxi stands at every 30-minute period.

TABLE III. MODELS ACCURACY USING ALPHA (α) = 0.4

MODEL	PERIODS			
	00->08	08->16	16->00	24h
Poisson Mean	74,99%	70,73%	72,60%	72,26%
W. Poisson Mean	75,73%	72,71%	74,40%	73,92%
ARIMA	77,11%	72,40%	74,87%	74,22%
Ensemble	**77,90%**	**74,73%**	**76,66%**	**76,04%**

TABLE IV. MODELS ACCURACY USING ALPHA (α) = 0.5

MODEL	PERIODS			
	00->08	08->16	16->00	24h
Poisson Mean	74,99%	70,73%	72,60%	72,26%
W. Poisson Mean	75,47%	71,45%	73,48%	72,98%
ARIMA	77,11%	72,40%	74,87%	74,22%
Ensemble	**78,07%**	**74,75%**	**76,62%**	**76,08%**

Our model demonstrated a more than satisfactory performance, predicting accurately more than 76% of the 86411 tested services, anticipating in real time the spatial distribution of the passenger demand. We believe that **this model is a true novelty and a major contribution** to the area by its online adapting characteristics (i.e. short term predictions) which can truly **improve the vehicles/drivers mobility intelligence** and consequently, their profit.

This model will be used as a feature of a recommendation system (to be done) which will produce smart live recommendations to the taxi driver about which taxi stand he should head to after a drop-off. We believe that the deployment of such system on a taxi fleet will contribute to increase its competitivity facing other taxi fleets in a Scenario 1 network (e.g. like the studied one, where the average waiting time to pick-up a passenger in a taxi stand is three times higher than the average service duration).

REFERENCES

[1] H. Yang, C. Cowina, S. Wong *et al.*, “Equilibria of bilateral taxi–customer searching and meeting on networks,” *Transportation Research Part B: Methodological,* vol. 44, no. 8–9, pp. 1067-1083, 2010.

[2] K. Wong, S. Wong, M. Bell *et al.*, “Modeling the bilateral micro-searching behavior for urban taxi services using the absorbing markov chain approach,” *Journal of Advanced Transportation,* vol. 39, no. 1, pp. 81-104, 2005.

[3] L. Bin, Z. Daqing, S. Lin *et al.*, “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, 2011, pp. 63-68.

[4] J. Lee, G.-L. Park, H. Kim *et al.*, "A Telematics Service System Based on the Linux Cluster," *Computational Science – ICCS 2007*, Lecture Notes in Computer Science, pp. 660-667: Springer Berlin / Heidelberg, 2007.

[5] L. Junghoon, S. Inhye, and G. Park, “Analysis of the Passenger Pick-Up Pattern for Taxi Location Recommendation,” in Conference on Networked Computing and Advanced Information Management, 2008, pp. 199-204.

[6] L. Liu, C. Andris, A. Biderman *et al.*, “Uncovering Taxi Driver’s Mobility Intelligence through His Trace,” *IEEE Pervasive Computing*, 2009.

[7] A. Ihler, Hutchins, J., Smyth, P., “Adaptive Event Detection with Time-Varying Poisson Processes,” in 12th ACM SIGKDD international conference on Knowledge discovery and data mining, Philadelphia, PA, USA, 2006, pp. 207-216.

[8] G. Box, and D. Pierce, “Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models,” *Journal of the American Statistical Association,* vol. 65, no. 332, pp. 1509-1526 1970.

[9] L. Matias, J. Gama, J. Mendes-Moreira *et al.*, “Validation of both number and coverage of bus Schedules using AVL data. ,” in ITSC'2010, Funchal, Portugal, 2010, pp. 131-136.

[10] C. Holt, “Forecasting seasonals and trends by exponentially weighted moving averages,” *International Journal of Forecasting,* vol. 20, no. 1, pp. 5-10, 2004.

[11] L. H. B.Williams, “Modeling and Forecasting Vehicular Traffic Flow as a Seasonal ARIMA Process: Theoretical Basis and Empirical Results,” *Journal of Transportation Engineering,* vol. 129, no. 6, pp. 664-672, 2003.

[12] G. P. Zhang, “Time series forecasting using a hybrid ARIMA and neural network model,” *Neurocomputing,* vol. 50, no. 0, pp. 159-175, 2003.

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

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

[15] H. Wang, W. Fan, P. S. Yu *et al.*, “Mining concept-drifting data streams using ensemble classifiers,” in Proceedings of the 9th ACM SIGKDD international conference on Knowledge discovery and data mining, Washington, D.C., 2003, pp. 226-235.

[16] S. Makridakis, and M. Hibon, “The M3-Competition: results, conclusions and implications,” *International Journal of Forecasting,* vol. 16, no. 4, pp. 451-476, 2000.

[17] M. Ferreira, H. Conceicao, R. Fernandes *et al.*, “Stereoscopic aerial photography: an alternative to model-based urban mobility approaches,” in Proceedings of the 6th ACM international workshop on VehiculAr InterNETworking, Beijing, China, 2009, pp. 53-62.

[18] R Development Core Team, *R: A Language and Environment for Statistical Computing.*, Vienna, Austria, 2005.

[19] K. Yeasmin, and J. H. Rob, "Automatic Time Series Forecasting: The forecast Package for R."