Predicting Taxi-Passenger Demand using Streaming Data

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

***Abstract*—*Informed driving* is becoming a key feature to increase the sustainability of taxi companies. The sensors installed in each vehicle are providing new opportunities to automatically discover knowledge, which in return deliver information for real-time decision making. Intelligent transportation systems for taxi dispatching and time-saving route finding are already exploring this sensing data. In this paper, we introduce a novel methodology to predict the spatial distribution of taxi-passenger in a short-term time horizon using streaming data. We have done so by firstly aggregating the information into a histogram time series. Then, we combined three time series forecasting techniques to output our prediction. Experimental tests were done using the online data transmitted by 441 vehicles of a fleet running in the city of Porto, Portugal. Our results demonstrated that the proposed framework can provide an effective insight into the spatiotemporal distribution of taxi-passenger demand in a 30 minutes horizon.**

***Index Terms*—taxi-passenger demand, mobility intelligence, GPS data, data streams,time series forecasting, auto-regressive integrated moving average (ARIMA), time-varying Poisson models, ensemble learning.**

I. Introduction

ADVANCES in sensor and wireless communications such as GPS (Global Positioning System), GSM (Global System for Mobile Communications) and WiFi have provided a new way to communicate with running vehicles whilst collecting relevant information about their status and location. The majority of taxi vehicles are now equipped with these kind of technologies, producing a new source of rich spatiotemporal information. Intelligent transportation systems for efficient taxi dispatching [1], time-saving route finding [2], [3], fuel-saving routing [4] and taxi-sharing [5] are already successfully exploring these kind of data and/or interfaces.

The rising cost of fuel has been decreasing the profit of both taxi companies and drivers. It causes an unbalanced relationship between passenger demand and the number of running taxis, thus decreasing the profits made by companies and also the passenger satisfaction levels [6]. S. Wong presented a relevant mathematical model to express this need for equilibrium in distinct contexts [7]. An equilibrium fault may lead to one of two scenarios: (Scenario 1) excess of vacant vehicles and excessive competition; (Scenario 2) larger waiting times for passengers and lower taxi reliability. However, a question remains open: Can we guarantee that the taxi spatial distribution over time will always meet the demand? Even when the number of running taxis already does?

The taxi driver mobility intelligence is an important factor to maximize both profit and reliability within every possible scenario. 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 be an advantage for the driver - especially when there is no economic viability of adopting random cruising strategies to find their next passenger. The GPS historical data is one of the main variables of this topic because it can reveal underlying running mobility patterns. Multiple works in the literature have already explored this type of data successfully with distinct applications such as smart driving [3], modeling the spatiotemporal structure of taxi services [8]–[10], building passenger-finding strategies [11], [12] or even predicting taxi location through a passenger-perspective [13] (in a Scenario 2 urban area). Despite their useful insights, the majority of the techniques reported are tested using offline test-beds, discarding some of the main advantages of this type of signal. In other words, they do not provide any live information about passenger location or the best route to pick-up one in this specific date/time while the GPS data is mainly a live data stream (i.e. a time ordered sequence of instances produced in real-time [14]).

In our work, we focus on the real-time 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). An intelligent approach regarding this problem will improve the network reliability for both companies and clients: a clever distribution of vehicles throughout stands will decrease the average waiting time to pick-up a passenger while the distance traveled will be more profitable. Passengers will also experience a lower waiting time to get a taxi (automatically dispatched or directly picked-up at a stand).

Manuscript submitted August 8, 2012; revised February XX, 2013.

This work was supported by the projects DRIVE-IN: "Distributed Routing and Infotainment through Vehicular Internet-working", MISC: "Massive Information Scavenging with Intelligent Transportation Systems", VTL: "Virtual Traffic Lights" and KDUS: "Knowledge Discovery from Ubiquitous Data Streams " under the Grants CMU-PT/NGN/0052/2008, MITPT/ITS-ITS/0059/2008, 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.

Luis Moreira-Matias and João Mendes-Moreira are with the LIAAD-INESC TEC and with DEI, Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias, s/n 4200-465 Porto - Portugal (phone: 00351-91-4221647; e-mail: luis.matias; jmoreira@fe.up.pt).

João Gama is with LIAAD-INESC TEC and Faculdade de Economia, Universidade do Porto, Rua Dr. Roberto Frias, s/n 4200-465 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, Lda., Avenida de França, 20, Sala 605, 4050-275 Porto – Portugal (e-mail: luis@geolink.pt).

On the other hand, this can present a true advantage for a fleet when facing other competitors.

The stand-choice problem is based on four key variables: the expected revenue for a service over time, the distance/cost relation with each stand, the number of taxis 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 focus on the passenger demand spatiotemporal complexity.

In this paper, **we present a model to predict the number of services that will emerge at a given taxi stand**. Specifically, it predicts the passenger demand over space (taxi stand) for a short-time horizon of P-minutes. This model reuses the information constantly transmitted/received by the telematics installed in each taxi about the current period to predict what will happen in the next one. Our goal is to predict at the instant t how many services will emerge during the future period $[t, t + P]$ at each existent taxi stand, reusing the real-time service count of $[t, t + P]$ to do the same for the instant $t + P$ and so on (i.e. the framework runs continuously in a stream). To do so, we adapted well-known time series forecasting techniques such as the time varying Poisson model [15] and ARIMA (AutoRegressive Integrated Moving Average) [16] to our problem. There are some works in the literature related to this problem, namely: 1) mining the best passenger-finding strategies [11], [12] and 2) dividing the urban area into attractive clusters based on the historical passenger demand (i.e.: city zones with distinct demand patterns) [8]–[10] predicting the passenger demand at certain urban hotspots [17]–[19]. The **major contribution** of this work facing this state-of-the-art is to build predictions about the spatiotemporal distribution of the taxi-passenger demand using streaming data. In fact, the reported works present offline test-beds while our framework was tested in an online environment.

As a case-study, we have selected a large-size taxi fleet running in the city of Porto, Portugal. The city contains a total of 63 taxi stands and two taxi companies running one fleet each. We used the data transmitted by the biggest one - which has 441 vehicles. In this network, each vehicle waits on average 44 minutes to pick-up a passenger (Scenario 1 city).

Our study just uses as input/output the services received directly at the stands or automatically dispatched to the parked vehicles, ignoring the remaining ones. This was done because the passenger demand at each taxi stand is the main feature to aid the taxi drivers' decision, since it represents 76% of the total number of services (note that calls to the taxi central are preferentially assigned to vehicles already parked at a taxi stand).

The test-bed ran continuously over a total of 9 months between August 2011 and April 2012. However, the model just produced predictions (i.e. it was stream-tested) in the last four. The results obtained demonstrated both efficiency and success: our framework had an aggregated error of just 23.97% using a predictive time horizon of just 30 minutes. The model used, in average, 38.12 seconds of processing time during our real-time test-bed. Such output clearly demonstrates that this model is an advance facing the existing state-of-art on predicting the spatiotemporal distribution of taxi-passenger demand in an urban area.

The remainder of the paper is structured as follows. Section 2 revises the existing literature regarding this topic. Section 3 formally presents our model. The fourth section describes how we acquired and preprocessed the dataset 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 experimental setup and metrics used to evaluate our model; then, the obtained results are detailed, followed by some important remarks about them. Finally, conclusions are drawn as well as our future work.

II. Literature Review

In the last decade, GPS-location systems have attracted the attention of both researchers and companies due to the new type of information available. Specifically, the ubiquitous characteristics of this location-aware sensors (i.e. portable; available everywhere) and of the information transmitted (i.e. a stream) increases the challenge. Moreover, they are usually tracking human behavior (individual or in group) and they can be used collaboratively to reveal their mobility patterns. Trains [20], Buses [21], [22] and Taxi Networks [17] are already successfully exploring these traces. Gonzalez *et. al* [23] uncovered the spatiotemporal regularity of human mobility, which were demonstrated in other activities such as electricity load [24] or freeway traffic flow [15], [25], [26].

Recently, multiple works have used the GPS historical data to analyze the spatial structure of the passenger demand. Deng *et. al* [8] mined this type of data to build and explore an origin-destination matrix in the city of Shanghai, China. Liu *et. al* [9] uses a 3D clustering technique to analyze the mobility intelligence spatial-patterns for both top and ordinary drivers. Yue *et. al* [10] discover the Level of Attractiveness (LOA) of urban-spatiotemporal clusters.

The works focused on passenger/taxi-finding strategies commonly use data from Scenario 2 cities, where the demand is largely superior to the supply. An innovative study was presented by Bin *et. al* [17]. Their goal was 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 in a large city in China. They made an empirical study on the impact of the selected features and its conclusions were validated by the feature selection tool. Lee *et. al* [12] constructed a framework to describe the spatiotemporal structure of the passenger demand on Jeju Island, South Korea. A customer-focused research was developed by Phithakkitnukoon *et. al* [13]: they aimed to predict where the vacant taxis will be over space and time to aid the clients in their daily scheduling and planning.

Ge *et. al* [27] provided a cost-efficient route recommendation model which was able to recommend sequences of pick-up locations. Their goal was to learn from the historical data transmitted from the most successful drivers to improve the profit of the remaining ones. Yuan *et. al* presented in [28] a very complete work containing methods about a) how to divide

the urban area into pick-up zones using spatial clustering; b) how a passenger can find a taxi; and c) which trajectory is the best to pick-up the next passenger. Although their results are promising, both approaches are focused on improving the trajectory of a single driver, discarding the current network status (i.e. the position of the remaining drivers).

Little works regarding the demand prediction problem exist. Kaltenbrunner *et. al* [18] detected the geographic and temporal mobility patterns over data acquired from a bicycle network running in Barcelona. It also directly addresses the prediction problem using an ARMA (AutoRegressive Moving Average) model. Their goal was to forecast the number of bicycles at a station to improve the stations spatial deployment. Chang *et. al* [19] presented a novel insight on demand prediction: they applied clustering to data extracted from large Asian cities. They used some key features besides location/time such as the weather. Their output was a hotness probability ratio over spatial clusters (i.e. real agglomeration of roads/streets) dependent on the driver location, discarding however the other taxis position.

In fact, the ARIMA models are well-known time series forecasting models by its short-term prediction performance [17]–[19], [26], [29]–[31] . The traffic flow short-term prediction is approached by Min *et. al* [26]: they use both historical data and spatial correlations between road segments to forecast the speed and the volume of the traffic within a road network. Despite the usefulness of their contribution, the spatial correlations are difficult to maintain/update in a real-time test-bed (their own is offline). The most similar work to our own is presented by Li *et. al* [17]. They present a recommendation system to improve the driver mobility intelligence. To do so, they used data from a taxi network running in Hangzhou, China (Scenario 2). Firstly, they calculated the city hotspots: urban areas where pick-ups occur more frequently. Secondly, they used ARIMA to forecast the pick-up quantity at these hotspots over periods of 60 minutes. Thirdly, they presented an improved ARIMA dependent both on time and daytype. Finally, they proposed a recommendation system based on the following variables: 1) the number of taxis already located at each hotspot; 2) the distance from the driver location to the hotspot in time and 3) the prediction about the number of services to be demanded in each one of them. Despite their good results, this approach has three weak points when compared against our own: 1) it just uses the most immediate historical data, discarding the mid and long-term memory of the system; 2) their test-bed uses minimum aggregation periods of 60 minutes over offline historical data (i.e. the next value prediction task on a time series goes easier as long as you increase its aggregation period) while we use short-term periods of 30 minutes; 3) the paper does not clearly describes how they update both the ARIMA model and the weights used by it.

All reported works (including the two last ones) have a common characteristic: they are tested using mainly historical data and their results were calculated using an offline test-bed. Our framework is a **short-term prediction model** which uses short, mid and long-term historical data as input. It reuses the real-time service count from each stand to calculate the demand for the next period. It was tested using an **online test-bed along a real time period** of nine months. The contribution of this work is to produce short-term predictions about the demand at a fixed point as a computational lightweight process without discarding the long-term system memory (i.e. historical data). To the best of our knowledge, such approach has no parallel in the literature. This model is formally presented in the following section.

III. The Model

This model is an extension of the one already presented in [32]. Let $S = \{s_1, s_2, ..., s_N\}$ be the set of N taxi stands of interest and $D = \{d_1, d_2, ..., d_j\}$ be a set of j possible passenger destinations. Our problem is to choose the best taxi stand at instant t according to our forecast about passenger demand distribution over the time stands for the period $[t, t + P]$. However, the present work (and model) is just focused on the prediction problem.

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. Our goal is to build a model which determines the set of service counts $X_{k,t+1}$ for the instant $t+1$ and per each taxi stand $k \in \{1, N\}$. 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 [15]. The demand for taxi services exhibits, like other modes of road transportation [21] , a periodicity in time on a daily basis that reflects the patterns of the underlying human activity, making the data appear non-homogeneous. Fig. 1 illustrates a one month taxi service analysis extracted from our dataset that illustrates this periodicity (the dataset is described in detail in Section IV).

Consider the probability to emerge n taxi assignments in a determined time period - $P(n)$ - following a **Poisson Distribution**. We can define it using the following equation

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

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 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

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

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 $h(t)$ the period in which time t falls (e.g. the time 00:31 is contained in period 2 if we consider 30-minutes periods).

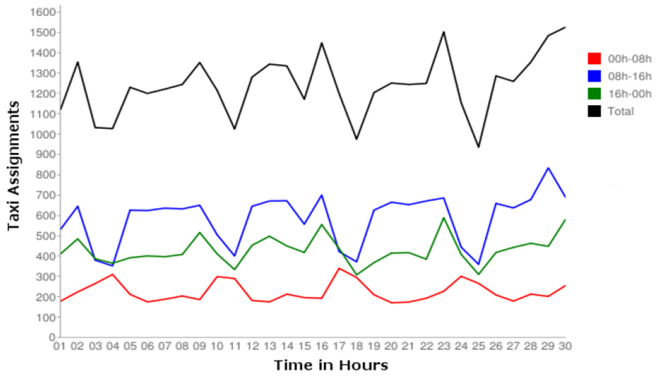

Fig. 1. One month data analysis (total and per shift).

Consider $\lambda(t)$ to be a discrete function (e.g.: an histogram time series of event counts aggregated in periods of P minutes). The equation (2) requires the validity of both equations

$$\sum_{i=1}^{7} \delta_i = 7 \tag{3}$$

$$\sum_{i=1}^{I} \eta_{d,i} = I, \forall d \tag{4}$$

where I is the number of time intervals in a day. As result, we have a discrete time series per stand representing the expected demand during an entire week: $\lambda(t)_k$. Each value of this series is an average of all the demands previously measured in the same daytype and period (i.e. the expected service demand for a Monday from 8:00 to 8:30 is the average of the demand on all past Mondays from the 8:00 to 8:30).

B. 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 of the seasonality demand : the taxi demand around them will be higher over summer weekends rather than over other seasons throughout the year.

To face this specific issue, we propose a weighted average model based on the one presented before: our goal is to increase the relevance of the demand pattern observed in the previous week by comparing it with the patterns observed several weeks ago (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 [33].

We can define ω as follows

$$\omega = \alpha * \{1, (1-\alpha), (1-\alpha)^2, ..., (1-\alpha)^{\gamma-1}\}, \gamma \in \mathbb{N} \tag{5}$$

where γ is the number of historical periods considered and $0 < \alpha < 1$ is the smoothing factor (i.e. γ and α are user-defined parameters). Then, based on the previous definition of $\lambda(t)_k$, we can define the resulting weighted average $\mu(t)_k$ as follows

$$\mu(t)_k = \sum_{i=1}^{\gamma} \frac{X_{t-(\theta * i)} * \omega_i}{\Omega}, \Omega = \sum_{i=1}^{\gamma} \omega_i \tag{6}$$

where θ represents the number of time periods contained in a week.

C. AutoRegressive Integrated Moving Average Model

The two previous models assume the existence of a regular (seasonal or not) periodicity in the 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 and to adapt to these changes instantly.

The AutoRegressive Integrated Moving Average Model (ARIMA) [16] is a well-known methodology to both model and forecast univariate time series data such as traffic flow data [26], electricity price [29] and other short-term prediction problems like our own. The ARIMA main advantages when compared to other algorithms are two: 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); 2) on the other hand, 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 enunciated below following the existing description in [30] (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 [31].

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 generates the time series (taxi service over time for a given stand k) as

$$\begin{aligned} R_{k,t} = \kappa_0 + \phi_1 X_{k,t-1} + \phi_1 X_{k,t-2} + ... + \phi_p X_{k,t-p} \\ +\varepsilon_{k,t} - \kappa_1 X_{k,t-1} - \kappa_1 X_{k,t-2} - ... - \kappa_q X_{k,t-q} \end{aligned} \tag{7}$$

where $R_{k,t}$ and $\varepsilon_{k,t}$ are the actual value and the random error at time period t, respectively; $\phi_l (l = 1, 2, ..., p)$ and $\kappa_m (m = 0, 1, 2, ..., q)$ are the model parameters/weights while p and q are positive integers often referred to 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 as has been proposed by Box and Jenkins in [16] They are useful to detect if the signal is periodic and, most important, which the frequencies of these periodicities are. A study conducted on time series from the demand of taxi services in one of the busiest taxi stands is displayed in Fig. 2.

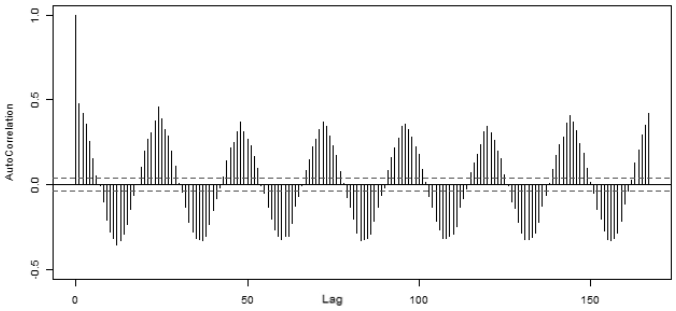

Fig. 2. Autocorrelation profile for data about the demand on taxi service (13 weeks) obtained from one of the busiest taxi stands in the city (periods of 60-minutes). The x-axis has the different period lags studied and the y-axis has the correlation within the signal. Note the peaks for each 12h periods.

D. Sliding Window Ensemble Framework

We already proposed three distinct predictive models which focused themselves to learn from the long, medium and short-term historical data. But a question remains open: how can we combine them all to improve our prediction? In the last decade, regression and classification tasks on streams attracted the community attention due to its drifting characteristics. The ensembles of such models were specifically addressed due to the challenge related with this type of data. One of the most popular models is the weighted ensemble [34]. The model we propose below is based on this one.

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

$$E_t = \sum_{i=1}^{z} \frac{F_{it}}{\Upsilon}, \Upsilon = \sum_{i=1}^{z} (1 - \rho_{iH}) \quad (8)$$

where ρ_{iH} is the error of the model M_i in the periods contained on the time window $[t - H, t]$ (H is a user-defined parameter to define the window size) while compared with the real service count time series. As the information is arriving in a continuous manner for the next periods $t, t+1, t+2, ...$ the window will also **slide** to determine how the models are performing in the **last H periods**.

To calculate such error, we used the Symmetric Mean Percentage Error (*sMAPE*), which is formally described in section V of this paper.

IV. Data Acquisition And Preprocessing

As a case-study, we focused on the stream event data of a taxi company operating in the city of Porto, Portugal. This city is the center of a medium size urban area (consisting of 1.3 million habitants) where the passenger demand is inferior to the number of running vacant taxis, resulting in a huge competition between both companies and drivers - according to a recent aerial survey of the road traffic of the city [35], taxis represent 4% of the running vehicles during a non-rush hour period. The existing regulations force the drivers to not run *randomly* in search of passengers but to choose a specific taxi stand out of the 63 existing ones in the city to wait for the next service immediately after the last passenger drop-off. A map of the stand spatial distribution is presented in the Fig. 3.

There are three main ways to pick-up a passenger: (1) a passenger goes to a taxi stand and picks-up a taxi – the regulations also force the passengers to pick-up the first taxi in line (First In, First Out); (2) a passenger calls the taxi network central and demands a taxi for a specific location/time – the parked taxis have priority over the running vacant ones in the central taxi dispatch system; (3) a passenger picks a vacant taxi while it is going to a taxi stand, on any street.

In this section, we describe the studied company, 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 of the company fleet. This taxi central usually runs 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 – relative to the type of event reported and 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 at 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) refers to the LATITUDE and the LONGITUDE corresponding to the acquired GPS position. This data was acquired over a non-stop period of nine months. Our 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). We did so because the passenger demand at each taxi stand 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 a period of P-minutes was developed. There are three types of accounted events: (1) the *busy* set directly at a taxi stand; (2) the *assign* set directly to a taxi parked at a taxi stand and (3) the *busy* set while a vacant taxi is cruising. We consider both a type 1 and type 2 event as service demanded.

Fig. 3. Taxi Stand spatial distribution over the city of Porto, Portugal.

However, for each type 2 event, the system receives a *busy* event a few minutes later – as soon as the driver effectively picked-up the passenger – this is ignored by our system. Type 3 events are ignored unless they occur in a radius of W meters from a taxi stand (where W is a user defined parameter). If it does, it is considered as being a type 1 event related with the nearest taxi stand according the defined criteria. This was done because many regulations prohibit the picking-up of passengers in a predefined radius of a stop (in Porto a 50m radius is in place). Some statistics about the studied period are now presented. Fig. 4 has the sample distribution of the cruise time of the services demanded. Table I details the number of taxi services demanded per daily shift and day type. Table II has information about all the services per taxi/driver and per cruise time. The *service* column in Table II represents the number of services picked-up by the taxi drivers, while the second one represents the total cruise time of every services 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 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 inability of the regulators to reduce the number of taxis in the city. It also highlights the importance of our recommendation system, where the shortness of services could be mitigated by getting services from the competitors.

The data in Tables I and II sustain that, despite the regularity exhibited in the service (especially on the weekends), there are big differences among the services performed per each driver

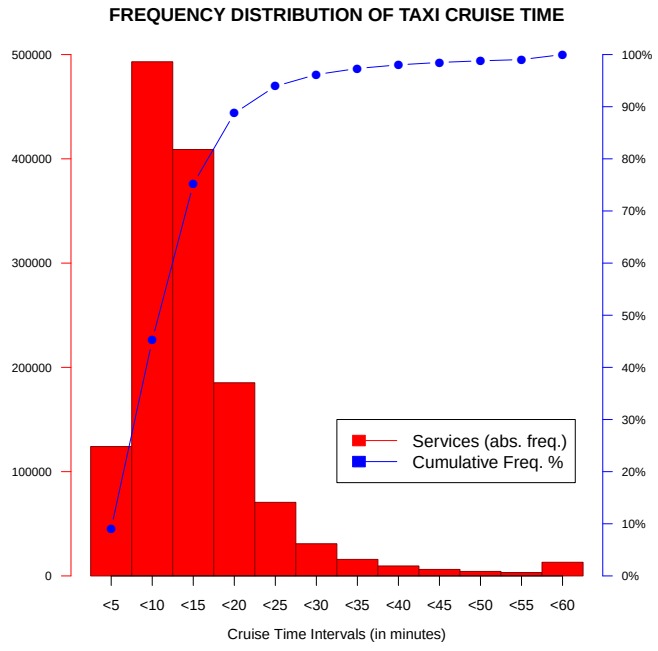

Fig. 4. Frequency Distribution of Taxi Cruise Time.

TABLE I
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 II
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

(i.e. a large variance in services number and profit) related with their distinct levels of mobility intelligence. Fig. 4 focuses on the length of the services: 75% of them last 15 minutes or less. These statistics sustain the importance of a smart decision on the stand-choice problem: an accurate sensor on the passenger demand can be a major advantage in urban areas where a highly competitive scenario – like our own - is in place.

V. EXPERIMENTAL RESULTS

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

A. Experimental Setup

Our test-bed was based on *prequential* evaluation [36]: data about the events occurring in the network was continuously acquired. We used an H-sized sliding window to measure the error of our model before each new prediction about the service count on the next period (the metrics used to do so are defined in the section V-B). Each new real count was used to update our predicting model.

Each data chunk was transmitted and received through a socket. The model was programmed using the R language [37]. 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. Fig. 5 illustrates the described test-bed: the $PP_i ... PP_t (t = 3)$ are the independent predicting processes – each one handle a predetermined group of taxi stands. 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 100m ($W = 100 > 50$ defined by the existing regulations). This aggregation 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

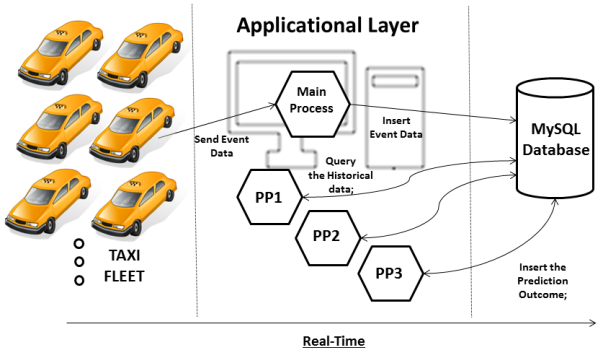

Fig. 5. Illustration about our streaming test-bed.

analysis) running on the historical time series curve of each stand during the last two weeks (i.e. period $t-2\theta, t$). To do so, we used an automatic time series function in the [forecast] R package [38] - *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.

A sensitivity analysis was conducted on parameter α based on a simplified version of Sequential Monte Carlo method (the reader can consult the survey in [39] to know more about this topic). Our goal was to calibrate our model by finding the optimal subregion on the input space $\alpha \in [0, 1]$ which maximizes our predictive performance. To do so, we generated 100 distinct samples as admissible values for α and we tested them using an older and smaller dataset containing data very similar to the one tested on our experiments (i.e. the same feature space). As result, we determined the ideal value as $\alpha = 0.4$. This value demonstrated to be robust - changes on it do not have a significant impact on the model output since they remain stable on the following input space: 0.4 ± 0.1. Therefore, we considered $\alpha = 0.4$ in our experiments. The γ value was set respecting the following definition

$$\gamma = \max(\mathbb{N}) : \omega_\gamma \geq 0.01 \tag{9}$$

which represents the limit for the weight $\omega_{i>\gamma} \sim 0$. According to this, $\alpha = 0.4 \implies \gamma = 8$.

Table III resumes the information about the learning periods used by each algorithm.

TABLE III
DESCRIPTION OF THE LEARNING PERIODS

Algorithm	Sliding Window	Nr. of Periods Considered
Poisson Mean	All Data $\{1, t\}$	N/A: it is calculated incrementally
W. Poisson Mean	Last two weeks	$\gamma = 8$
ARIMA	Last two weeks	$2 * \theta$
Ensemble	Last four hours	$H = 8$

B. Evaluation Metrics

We used the data obtained from the last four months to evaluate our framework (where 506873 services emerged). A well-known error measurement was employed to evaluate our output: the Symmetric Mean Percentage Error (*sMAPE*) [40]. We formally define it below.

Consider $R = \{R_{k,1}, R_{k,2}, ..., R_{k,t}\}$ 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 $X = \{X_{k,1}, X_{k,2}, ..., X_{k,t}\}$ the number of services actually emerged in the same conditions. The $(sMAPE_k)$ (i.e.: the error measured on the time series of services predicted to the stand k) can be defined as

$$sMAPE_k = \frac{1}{t}\sum_{i=1}^{t}\frac{|R_{k,i} - X_{k,i}|}{\varrho_{k,i}} \tag{10}$$

$$\varrho_{k,i} = \begin{cases} R_{k,i} + X_{k,i} & \text{if } (R_{k,i} > 0 \vee X_{k,i} > 0) \\ 1 & \text{if } (R_{k,i} = 0 \wedge X_{k,i} = 0) \end{cases} \tag{11}$$

where t is the number of time periods considered. 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 measured during that period would be 1). To produce more accurate statistics about series containing very small numbers, we can add a Laplace estimator [41] to (10). In this case, we will do it by adding a constant c to the denominator (i.e.: originally, it was added to the numerator to estimate a success rate [41]). Therefore, we can re-define $sMAPE_k$ as follows

$$sMAPE_k = \frac{1}{t}\sum_{i=1}^{t}\frac{|R_{k,i} - X_{k,i}|}{R_{k,i} + X_{k,i} + c} \tag{12}$$

where c is a user-defined constant. To simplify the theorem application, we will consider its most common use: $c = 1$ [41].

This metric is focused just on one time series for a given taxi stand k. However, the results presented below use an averaged error measure based on all stands series – AG. Consider β to be an error metric of interest. $AG_{\beta,t}$ is an aggregated metric given by a weighted average of the error measured in all stands in the period $1, t$. It is formally presented in the following equations:

$$AG_{\beta,t} = \sum_{k=1}^{N}\frac{\beta_{t,k} * \psi_k}{\Psi} \tag{13}$$

$$\psi_k = \sum_{i=1}^{t} X_{k,i}, \Psi = \sum_{k=1}^{N}\psi_k \tag{14}$$

where ψ_k is the total of services emerged at the taxi stand k; $\beta_{t,k}$ is the error measured by β at the stand k and Ψ is the total of services emerged at all stands so far.

C. Results

The results are presented over four distinct perspectives: 1) averaged error of the proposed methods; 2) a comparative analysis of the ensemble performance versus the remaining models; 3) a direct analysis of some output examples and 4) a small report about the computational time needed to predict the next period.

Firstly, the error measured for each model is presented in Table IV. The results are firstly presented per shift and then globally. The results were aggregated using the AG_β previously defined.

Secondly, Fig. 6 presents a comparison between our Ensemble and the other predictive models on a typical workday. These values were calculated using the same 4-hours sliding window of the ensemble (the error of the instant t is the error measured at the period $[t-H, t], H = 8$).

Thirdly, three distinct weekly analysis of the discrepancies between the demand predicted and the services actually emerged are displayed in Fig. 7. Our model forecasted the spatiotemporal taxi-passenger demand for every 30-minute period using (on average) 38.12 sec. of processing time (i.e. 1.906 sec. per time series/stand) as result of the computational parallel approach presented before. This method reduced the computational time by 70% (i.e. in the first three weeks, we tested our model using just one iterative process – one program, one CPU core – and it lasted, on average, 99.77 seconds). The ARIMA model update was also fast: 48.12 seconds (mean value). These results are discussed next.

D. Discussion

The overall performance is very good: the maximum value of the error was 28.23%. The sliding window ensemble is always the best model in every shift and period considered: the error measured was always lower than 26%. The models just present slight discrepancies within the daily shifts.

Our ensemble methodology is robust when compared with the remaining models: in Fig. 6 it is possible to identify a point where the ensemble maintained its performance while two other methods had a huge drop, highlighting the inherited learning of the ensemble approach. Fig. 7 presents two distinct scenarios to compare the demand forecasted with the real one: in A), the demand corresponds to an irregular taxi stand where services do not have an usual pattern to emerge (even if the demand is low); in B) the chart corresponds to a completely regular stand behavior. The two examples illustrate that our ensemble can actually correctly forecast the demand in distinct scenarios, periods and time horizons.

In our scenario, the target variable is the number of services to arise along a taxi stand network during a pre-defined period of time. It was chosen due to the stand relevance in this scenario (where 76% of the total number of services is directly demanded on them). However, this is not the reality in many big cities around the world due to their (de)regulation [6]. Most of the literature about this topic divide their scenarios/urban areas into spatial clusters - as exemplified in Fig. 8 - to predict and/or characterize the pick-up quantity distribution on a short-term time horizon [8]–[10], [17], [19], [27], [28]. Our mathematical model does not depend on how the services historical data is spatially aggregated (i.e. by stand or by spatial cluster) but only on the aggregation period of P-minutes (which is user-defined). Therefore, it also represents a straightforward contribution to previous work.

TABLE IV
ERROR MEASURED ON THE MODELS USING $sMAPE$

Model	Periods			
	00h−08h	08h−16h	16h−00h	24h
Poisson Mean	27.54%	24.00%	24.87%	25.09%
W. Poisson Mean	26.48%	24.34%	25.18%	24.84%
ARIMA	28.23%	24.70%	24.93%	27.00%
Ensemble	**25.85%**	**23.12%**	**23.89%**	**23.97%**

VI. Conclusions and Future Work

In this paper, we present 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 from a company operating in Porto, Portugal (where the passenger demand is lower than the number of vacant taxis) into time series of interest to use firstly (1) as learning base to our model and secondly (2) as a streaming test framework. As a result, our model was able to predict the taxi-passenger demand at each one of the 63 taxi stands at 30-minute period intervals.

Our model demonstrated a more than satisfactory performance, correctly predicting the 506873 tested services with an aggregated error measure lower than 26%. We believe that **this model is a true novelty and a major contribution** to the area through its adapting characteristics:

- It mines both the periodicity and seasonality of the passenger demand, updating itself regularly;
- It simultaneously uses long-term, mid-term and short term historical data as a learning base;
- It takes advantage of the ubiquitous characteristics of a taxi network, assembling the experience and the knowledge of all vehicles/drivers while they usually use just their own;

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.

This model will be used as a feature for a recommendation system (to be done) which will produce smart live recommendations to the taxi driver about which taxi stand he should head

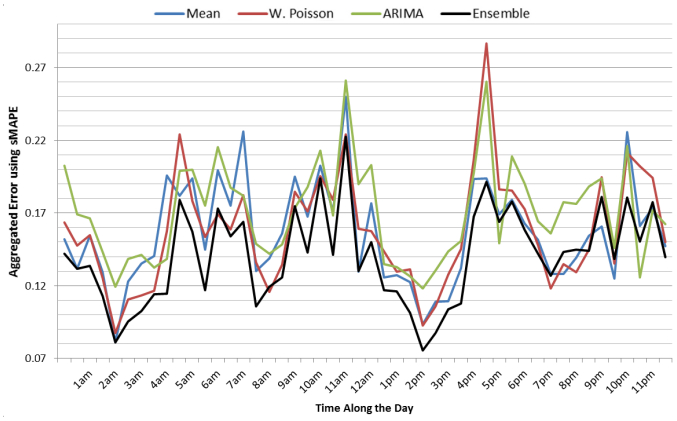

Fig. 6. Ensemble evaluation on a typical Saturday.

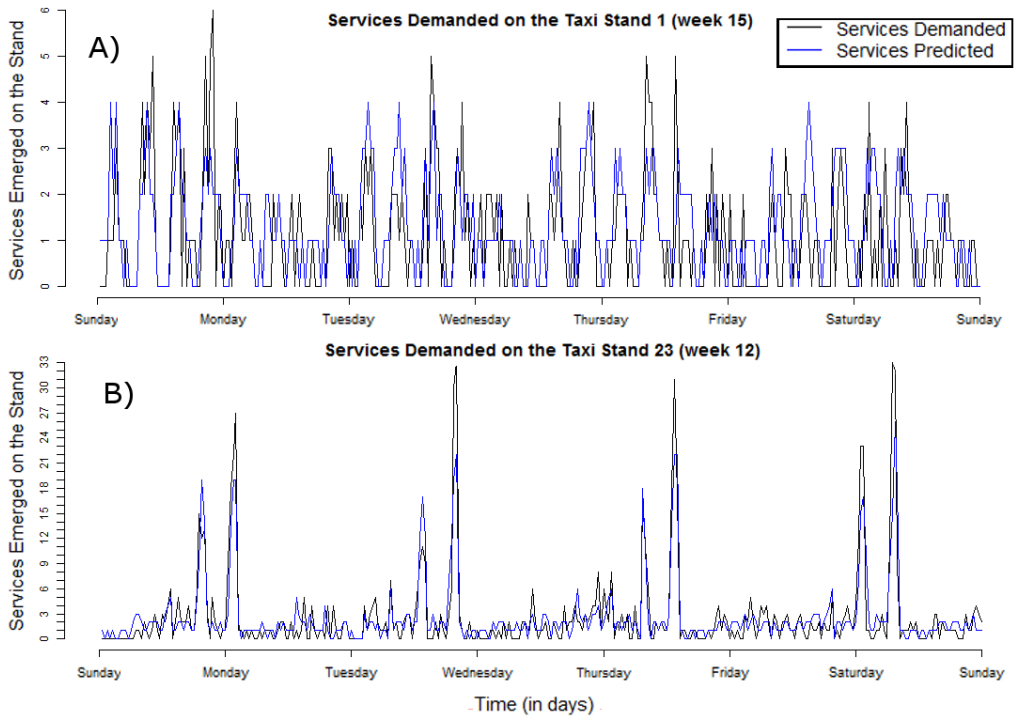

Fig. 7. Weekly comparison between the services forecasted and the services emerged on two distinct scenarios / taxi stands and weeks.

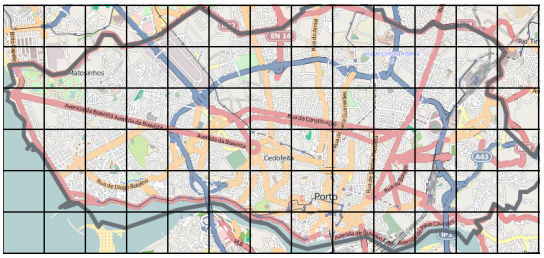

Fig. 8. Example of a possible spatial clustering of the city of Porto, Portugal.

to after a drop-off. This decision support framework will also address other features such as the distance or the live traffic conditions, among others. We believe that the deployment of such a system in 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 at a taxi stand is three times higher than the average service duration) by improving the distribution of the vacant vehicles throughout the stands.

Acknowledgments

The authors would like to thank Geolink, Lda. and to its team for the data supplied. We also thank the anonymous reviewers for their valuable comments and suggestions to improve this work.

References

[1] A. Glaschenko, A. Ivaschenko, G. Rzevski, and P. Skobelev, "Multi-agent real time scheduling system for taxi companies," in *8th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2009), Budapest, Hungary*, 2009, pp. 29–36.

[2] J. Lee, G. Park, H. Kim, Y. Yang, P. Kim, and S. Kim, *A Telematics Service System Based on the Linux Cluster*, ser. LNCS. Springer Berlin / Heidelberg, 2007, vol. 4490, pp. 660–667.

[3] 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*. ACM, 2010, pp. 99–108.

[4] P. Chen, J. Liu, and W. Chen, "A fuel-saving and pollution-reducing dynamic taxi-sharing protocol in vanets," in *2010 IEEE 72nd Vehicular Technology Conference Fall (VTC 2010-Fall)*. IEEE, 2010, pp. 1–5.

[5] P. d'Orey, R. Fernandes, and M. Ferreira, "Empirical evaluation of a dynamic and distributed taxi-sharing system," in *2012 15th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, Sept. 2012, pp. 140–146.

[6] B. Schaller, "Entry controls in taxi regulation: Implications of us and canadian experience for taxi regulation and deregulation," *Transport Policy*, vol. 14, no. 6, pp. 490–506, 2007.

[7] K. Wong, S. Wong, M. Bell, and H. Yang, "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.

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

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

[10] 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*. IEEE, 2009, pp. 1–6.

[11] 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)*, March 2011, pp. 63–68.

[12] J. Lee, I. Shin, and G. 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)*, vol. 1. IEEE, 2008, pp. 199–204.

[13] S. Phithakkitnukoon, M. Veloso, C. Bento, A. Biderman, and C. Ratti,

"Taxi-aware map: identifying and predicting vacant taxis in the city," *Ambient Intelligence*, vol. 6439, pp. 86–95, 2010.
[14] J. Gama, *Knowledge discovery from data streams*. Chapman and Hall/CRC, 2010.
[15] A. Ihler, J. Hutchins, and P. Smyth, "Adaptive event detection with time-varying poisson processes," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2006, pp. 207–216.
[16] G. Box, G. Jenkins, and G. Reinsel, *Time series analysis*. Holden-day San Francisco, 1976.
[17] X. Li, G. Pan, Z. Wu, G. Qi, S. Li, D. Zhang, W. Zhang, and Z. Wang, "Prediction of urban human mobility using large-scale taxi traces and its applications," *Frontiers of Computer Science in China*, vol. 6, no. 1, pp. 111–121, 2012.
[18] A. Kaltenbrunner, R. Meza, J. Grivolla, J. Codina, and R. Banchs, "Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system," *Pervasive and Mobile Computing*, vol. 6, no. 4, pp. 455–466, 2010.
[19] H. Chang, Y. Tai, and J. Hsu, "Context-aware taxi demand hotspots prediction," *International Journal of Business Intelligence and Data Mining*, vol. 5, no. 1, pp. 3–18, 2010.
[20] B. Cule, B. Goethals, S. Tassenoy, and S. Verboven, "Mining train delays," in *Advances in Intelligent Data Analysis X*, ser. LNCS, vol. 7014. Springer Berlin / Heidelberg, 2011, pp. 113–124.
[21] L. Matias, J. Gama, J. Mendes-Moreira, and J. Freire de Sousa, "Validation of both number and coverage of bus schedules using avl data," in *13th IEEE Conference on Intelligent Transportation Systems (ITSC)*, 2010, pp. 131–136.
[22] L. Moreira-Matias, C. Ferreira, J. Gama, J. Mendes-Moreira, and J. de Sousa, "Bus bunching detection by mining sequences of headway deviations," in *Advances in Data Mining. Applications and Theoretical Aspects*, ser. LNCS, vol. 7377. Springer., 2012, pp. 77–91.
[23] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, 2008, 10.1038/nature06958.
[24] J. Gama and P. Rodrigues, "Stream-based electricity load forecast," in *Knowledge Discovery in Databases: PKDD 2007*, ser. LNCS, vol. 4702. Springer Berlin / Heidelberg, 2007, pp. 446–453.
[25] B. Williams and L. Hoel, "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.
[26] W. Min and L. Wynter, "Real-time road traffic prediction with spatio-temporal correlations," *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 4, pp. 606–616, 2011.
[27] Y. Ge, H. Xiong, A. Tuzhilin, K. Xiao, M. Gruteser, and M. Pazzani, "An energy-efficient mobile recommender system," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2010, pp. 899–908.
[28] J. Yuan, Y. Zheng, L. Zhang, X. Xie, and G. Sun, "Where to find my next passenger?" in *13th ACM International Conference on Ubiquitous Computing, UbiComp 2011*, 2011.
[29] J. Contreras, R. Espinola, F. J. Nogales, and A. J. Conejo, "Arima models to predict next-day electricity prices," *IEEE Transactions on Power Systems*, vol. 18, no. 3, pp. 1014–1020, 2003.
[30] G. Zhang, "Time series forecasting using a hybrid arima and neural network model," *Neurocomputing*, vol. 50, pp. 159–175, 2003.
[31] J. Cryer and K. Chan, *Time Series Analysis with Applications in R*. USA: Springer, 2008.
[32] L. Moreira-Matias, J. Gama, M. Ferreira, and L. Damas, "A predictive model for the passenger demand on a taxi network," in *Intelligent Transportation Systems (ITSC), 2012 15th International IEEE Conference on*, sept. 2012, pp. 1014 –1019.
[33] C. Holt, "Forecasting seasonals and trends by exponentially weighted moving averages," *International Journal of Forecasting*, vol. 20, no. 1, pp. 5–10, 2004.
[34] H. Wang, W. Fan, P. 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*. ACM, 2003, pp. 226–235.
[35] M. Ferreira, H. Conceição, R. Fernandes, and O. Tonguz, "Stereoscopic aerial photography: an alternative to model-based urban mobility approaches," in *Proceedings of the sixth ACM international workshop on VehiculAr InterNETworking*. ACM, 2009, pp. 53–62.
[36] A. Dawid, "Present position and potential developments: Some personal views: Statistical theory: The prequential approach," *Journal of the Royal Statistical Society. Series A (General)*, vol. 147, pp. 278–292, 1984.
[37] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2012. [Online]. Available: http://www.R-project.org
[38] K. Yeasmin and J. H. Rob, *Automatic Time Series Forecasting: The forecast Package for R*, 1999. [Online]. Available: http://oai.repec.openlib.org
[39] O. Cappé, S. Godsill, and E. Moulines, "An overview of existing methods and recent advances in sequential monte carlo," *Proceedings of the IEEE*, vol. 95, no. 5, pp. 899–924, 2007.
[40] S. Makridakis and M. Hibon, "The m3-competition: results, conclusions and implications," *International Journal of Forecasting*, vol. 16, no. 4, pp. 451–476, 2000.
[41] E. Jaynes, *Probability theory: The logic of science*. Cambridge university press, 2003.

Luis Moreira-Matias received his Master's degree in Informatics Engineering from the Faculty of Engineering of the University of Porto (FEUP), Portugal, in 2009. Currently, he is a PhD student on Machine Learning at FEUP. His current research interests include learning from data streams, Intelligent Transportation Systems and time series forecasting problems.

João Gama is a researcher at LIAAD, the Laboratory of Artificial Intelligence and Decision Support of the University of Porto, working at the Machine Learning group. His main research interest is in Learning from Data Streams. He is the author of a recent book on Knowledge Discovery from Data Streams.

Michel Ferreira received the B.Sc. degree in computer science and the Ph.D. degree in computer science from the Universidade do Porto, Porto, Portugal, in 1994 and 2002, respectively. Currently, he is Assistant Professor on the FCUP - Faculty of Sciences of the University of Porto. He leads the Geo-Networks group on the Department of Computer Science - FCUP. He has led several research projects in the areas of logic-based spatial databases, vehicular sensing, and intervehicle communication. He is also co-founder of Geolink, Lda.

João Mendes-Moreira received his PhD degree in Engineering Sciences from the Faculty of Engineering of the Porto University (FEUP). He is an Assistant Professor in the Department of Informatics Engineering, FEUP, and researcher at LIAAD (Laboratory for Artificial Intelligence and Decision Aid), a group belonging to INESC Porto LA. He has worked in projects and authored papers in areas that are mainly related to the application of machine learning to problems of transport planning.

Luis Damas received the the Ph.D. degree in Mathematical Theory of Computation from University of Edinburgh in 1984. He is the creator of one of the most widely used logic programming systems - the YAP Prolog compiler; and of the DIVERT traffic simulator. He is also co-founder and CTO of Geolink, Lda, a spin-off from University of Porto, specialized in vehicular dispatching systems. His research interests include vehicular networks, distributed systems, simulation, programming languages and theory of computation.