Predicting Taxi–Passenger Demand Using Streaming Data

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

***Abstract*—*Informed driving* is increasingly becoming a key feature for increasing the sustainability of taxi companies. The sensors that are installed in each vehicle are providing new opportunities for automatically discovering knowledge, which, in return, delivers information for real-time decision making. Intelligent transportation systems for taxi dispatching and for finding time-saving routes are already exploring these sensing data. This paper introduces a novel methodology for predicting the spatial distribution of taxi–passengers for a short-term time horizon using streaming data. First, the information was aggregated into a histogram time series. Then, three time-series forecasting techniques were combined to originate a prediction. Experimental tests were conducted using the online data that are transmitted by 441 vehicles of a fleet running in the city of Porto, Portugal. The results demonstrated that the proposed framework can provide effective insight into the spatiotemporal distribution of taxi–passenger demand for a 30-min horizon.**

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

I. Introduction

ADVANCES in sensor and wireless communications such as Global Positioning System (GPS), Global System for Mobile Communications (GSM), and WiFi have provided a new way of communicating with running vehicles while collecting relevant information on their status and location. Most taxi vehicles are now equipped with these kinds 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 reducing the profit for both taxi companies and drivers. This leads to an unbalanced relationship between the passenger demand and the number of running taxis, which, in turn, reduces the companies' profits and also the levels of passenger satisfaction [6]. Yang *et al.* presented a relevant mathematical model for expressing this need for an equilibrium in distinct contexts [7]. A failure in this equilibrium may lead to one of the following two scenarios: Scenario 1, i.e., an excess in vacant vehicles and competition and Scenario 2, i.e., larger waiting times for passengers and lower taxi reliability. However, a question remains open. Is it possible to guarantee that the taxis' spatial distribution over time will always meet the demand, even when the number of running taxis already does that?

Taxi-driver mobility intelligence is an important factor in maximizing both profit and reliability within every possible scenario. Knowledge on where the services (transporting a passenger from pick-up to drop-off locations) will actually emerge can be an advantage for the driver, particularly when there is no economic viability of adopting random cruising strategies to find passengers. GPS historical data are one of the main variables of this topic because it can reveal underlying running mobility patterns. Multiple works in the literature have already successfully explored this type of data on various 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, most reported techniques are tested using offline testbeds, discarding some of the main advantages of this type of signal. In other words, they do not provide any live information on the location of a passenger or the best route to pick up a passenger at the *current* specific date/time (i.e., real-time performance), while the GPS data are mainly a live data stream (i.e., a time-ordered sequence of instances that are produced in real-time [14]).

This paper focuses on the real-time choice problem of which is the best taxi stand to go to after a passenger drop-off (i.e., the stand where another passenger can be picked up more quickly).

Manuscript received August 7, 2012; revised February 22, 2013 and April 20, 2013; accepted April 26, 2013. Date of publication June 14, 2013; date of current version August 28, 2013. This work was supported in part by Projects "Distributed Routing and Infotainment through Vehicular Internet-working" (DRIVE-IN), "Massive Information Scavenging with Intelligent Transportation Systems" (MISC), "Virtual Traffic Lights" (VTL), and "Knowledge Discovery from Ubiquitous Data Streams" (KDUS) under Grant CMU-PT/NGN/0052/2008, Grant MITPT/ITS-ITS/0059/2008, Grant PTDC/EIA-CCO/118114/2010, and Grant PTDC/EIA-EIA/098355/2008, respectively; by the European Regional Development Fund (ERDF) through the COMPETE Programme (Operational Programme for Competitiveness), and by the Portuguese Funds through the Portuguese Foundation for Science and Technology (FCT) within Project FCOMP-01-0124-FEDER-022701. The Associate Editor for this paper was Dr. M. Brackstone.

L. Moreira-Matias and J. Mendes-Moreira are with the Laboratory for Artificial Intelligence and Decision Support, Tecnologia e Ciência, Instituto de Engenharia de Sistemas e Computadores, and also with the Department of Informatics Engineering, Faculdade de Engenharia, Universidade do Porto, Porto 4200-465, Portugal (e-mail: luis.matias@fe.up.pt; jmoreira@fe.up.pt).

J. Gama is with the Laboratory for Artificial Intelligence and Decision Support, Tecnologia e Ciência, Instituto de Engenharia de Sistemas e Computadores, and also with the Faculdade de Economia, Universidade do Porto, Porto 4200-465, Portugal (e-mail: jgama@fep.up.pt).

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

L. Damas is with Geolink, Lda., Porto 4050-275, Portugal (e-mail: luis@geolink.pt).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2013.2262376

An intelligent approach regarding this problem will improve network reliability for both companies and clients; an intelligent distribution of vehicles throughout stands will reduce the average waiting time to pick up a passenger, while the distance traveled will be more profitable. Furthermore, whenever they need a taxi, passengers will also experience a lower waiting time to get a vacant taxi (automatically dispatched or directly picked up at a stand). This is competitively a true advantage for a fleet versus its competitors.

The stand-choice problem is based on four key variables: 1) the expected revenue for a service over time; 2) the distance/cost relation with each stand; 3) the number of taxis that are already waiting at each stand; and 4) the passenger demand for each stand over time. The taxi vehicular network can be a ubiquitous sensor of taxi–passenger demand from where the aforementioned variables can be continuously mined. However, the work described here will just focus on the spatiotemporal complexity of the passenger demand.

This paper **presents a model for predicting 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 that is constantly transmitted/received by the telematics that are installed in each taxi about the current period to predict what will happen in the next one. The aim is to predict at the instant (t) how many services will emerge during the future period $([t, t+P])$ at each existing taxi stand. The same will be performed for the following instant $(t+P)$ by reusing the real-time service count of $([t, t+P])$, and so on (i.e., the framework continuously runs in a stream). To do so, well-known time-series forecasting techniques were used and adapted to this problem, such as the time-varying Poisson model [15] and the autoregressive integrated moving average (ARIMA) [16]. There are works in the literature that are related to this problem, namely, mining the best passenger-finding strategies [11], [12] and dividing the urban area into attractive clusters based on historical passenger demand (i.e., city zones with distinct demand patterns) [8]–[10], thus making it possible to predict the passenger demand at certain urban hotspots [17]–[19]. The **major contribution** of this paper facing this state of the art is to build predictions on the spatiotemporal distribution of the taxi–passenger demand using streaming data. The existing research presents offline testbeds, whereas the framework that is presented here was tested in an online environment.

A large-size taxi fleet running in the city of Porto, Portugal, was selected as a case study. The city contains a total of 63 taxi stands and two taxi companies, each of which is running one fleet. The data that are transmitted by the largest company, which has 441 vehicles, was used. In this network, each vehicle waits for 44 min on average to pick up a passenger (Scenario-1 city).

The study presented here uses as input/output the services directly received 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 that is aiding the taxi drivers' decisions, since it represents 76% of the total number of services (note that calls to the taxi central are preferentially assigned to vehicles that are already parked at a taxi stand).

The testbed continuously ran over a total of nine months between August 2011 and April 2012. However, the model simply produced predictions (i.e., it was stream tested) for the last four months. The results that are obtained were both efficient and successful; the framework presented an aggregated error of just 23.97% using a predictive time horizon of 30 min. On average, the model used 38.12 s of processing time during our real-time testbed. Such output clearly demonstrates that this model is comparatively an advance to the existing state of the art on predicting the spatiotemporal distribution of the taxi–passenger demand in an urban area.

The remainder of this paper is structured as follows. Section II revises the existing literature on this topic. Section III formally presents the model employed. Section IV first describes how the data set used was acquired and preprocessed. Then, some statistics about it are presented. Section V describes how the methodology was tested in a real scenario. First, the experimental setup and metrics that are used to evaluate the model are introduced, and then, the results obtained are presented in detail, which is followed by some important remarks. Finally, conclusions are drawn and future work topics are discussed.

II. Literature Review

Over the last decade, GPS-location systems have attracted the attention of both researchers and companies due to the new type of information they provide. More specifically, the ubiquitous characteristics of these location-aware sensors and of the information transmitted (i.e., a stream) make these systems increasingly challenging. Moreover, these sensors usually track human behavior (individual or in group), and they can be collaboratively used to reveal 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 was demonstrated in other activities such as electricity load [24] or freeway traffic flow [15], [25], [26].

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

Research works that are focused on passenger/taxi-finding strategies commonly use data from Scenario-2 cities, where the demand largely exceeds the supply. An innovative study was presented by Li *et al.* [11]. Their goal was to validate the triplet time–location–strategy as the key features to build a good passenger-finding strategy. They used an L1-norm support vector machine as a feature selection tool to discover both efficient and inefficient passenger-finding strategies in a large city in China. They conducted an empirical study on the impact of the selected features, and their conclusions were validated by the feature selection tool. Lee *et al.* [12] created a framework to

describe the spatiotemporal structure of the passenger demand on Jeju Island, South Korea. A customer-focused approach was developed by Phithakkitnukoon *et al.* [13], i.e., to predict where and when the vacant taxis will be 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 data that are transmitted from the most successful drivers to improve the profit of the remaining ones. Yuan *et al.* presented in [28] a complete work containing methods about the following: 1) how to divide the urban area into pick-up zones using spatial clustering; 2) how a passenger can find a taxi; and 3) 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, disregarding the position of the remaining drivers.

Little research regarding the demand prediction problem exists. Kaltenbrunner *et al.* [18] detected the geographic and temporal mobility patterns over data that are acquired from a bicycle network running in Barcelona. This paper also addresses the prediction problem using an autoregressive moving average (ARMA) model. The authors' 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; the authors applied clustering to the data that are extracted from large Asian cities, using other key features aside from location/time such as the weather. Their output was a hotness probability ratio over spatial clusters (i.e., real agglomeration of roads/streets) depending on the driver's location. However, the authors disregard the position of other taxis.

ARIMA models are time-series forecasting models that are widely known for their short-term prediction performance [17]–[19], [26], [29]–[31]. The short-term prediction of traffic flow is addressed by Min and Wynter [26]. The authors use both historical data and spatial correlations between road segments to forecast the speed and the volume of traffic in a road network. Although their contribution is useful, the spatial correlations are difficult to maintain/update in a real-time testbed (their testbed was performed offline). The most similar work to our own is presented by Li *et al.* [17]. The authors present a recommendation system for improving the drivers' mobility intelligence. To do so, data from a taxi network running in Hangzhou, China (Scenario 2), was used. First, they calculated the city hotspots, i.e., urban areas where pick-ups more frequently occur. Second, they used ARIMA to forecast the amount of pick-ups at these hotspots over periods of 60 min. Third, they presented an improved ARIMA depending both on time and day type. Finally, they proposed a recommendation system based on the following variables: 1) the number of taxis that are already located at each hotspot; 2) the distance from the driver' location to the hotspot in terms of time; and 3) the prediction of the number of services to be demanded in each one of them. Despite their good results, this approach comparatively has the following three weak points to the one presented: 1) it just uses the most immediate historical data, discarding the mid- and long-term memory of the system; 2) in their testbed, the authors use minimum aggregation periods of 60 min over offline historical data (i.e., the next value prediction task on a time series is easier as long as the aggregation period is increased), whereas we use short-term periods of 30 min; and 3) the work does not clearly describe how the authors update both the ARIMA model and the weights that are used by it.

All aforementioned research works (including the last two approaches) have a common characteristic, as mainly historical data were used and their results were calculated using an offline testbed. The framework presented here is a **short-term prediction model**, which uses short-, mid-, and long-term historical data as input. It reuses the number of services in real time from each stand to calculate the demand for the following period. It was tested using an **online testbed along a real-time period** of nine months. The main contribution of this paper is that it produces short-term predictions for the demand at a fixed point. This is a computational lightweight process that does not disregard the long-term system memory. To the best of our knowledge, such an approach has no parallel in the literature. The model is presented in the following section.

III. Model

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

Consider $X_k = \{X_{k,0}, X_{k,1}, \ldots, X_{k,t}\}$ 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 that determines the set of service counts $X_{k,t+1}$ for instant $t + 1$ and per taxi stand $k \in \{1, \ldots, N\}$. 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. These models will be described next.

A. Time-Varying Poisson Model

The following section presents a model that was first proposed in [15]. The demand for taxi services exhibits, like other modes of road transportation [21], a daily periodicity that reflects the patterns of the human activity. As a result, the data appear to be nonhomogeneous. Fig. 1 shows a one-month taxi-service analysis extracted from our data set that illustrates this periodicity (the data set is described in detail in Section IV).

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:

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

where λ represents the rate (average demand for taxi services) in a fixed time interval. However, in this specific problem, rate λ is not constant but time variant. Therefore, it was adapted as a function of time, i.e., $\lambda(t)$, transforming the Poisson

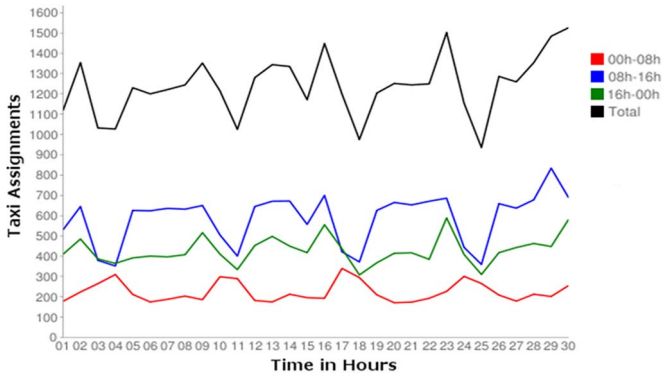

Fig. 1. One-month data analysis (total and per shift).

distribution into a nonhomogeneous 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 weekday $d(t)$ (e.g., Saturdays have lower day rates than Tuesdays); $\eta_{d(t),h(t)}$ is the relative change for period $h(t)$ in day $d(t)$ (e.g., the peak hours); $d(t)$ represents weekday $1 = \text{Sunday}, 2 = \text{Monday}, \ldots$; and $h(t)$ represents the period when time t falls (e.g., the time of 00:31 is contained in period 2 if we consider 30-min periods).

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

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

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

where I is the number of time intervals in a day. The result is a discrete time series per stand representing the expected demand during an entire week, i.e., $\lambda(t)_k$. Each value in this series is an average of all demands that are previously measured in the same day type 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 8:00 to 8:30).

B. Weighted Time-Varying Poisson Model

The model that is previously presented can be seen as a time-dependent average, which produces predictions based on long-term historical data. However, it is not guaranteed that every taxi stand will have a highly regular passenger demand; in fact, the demand in many stands can be often **seasonal**. The beaches are a good example of the seasonality demand as taxi demand will be higher during summer weekends as opposed to other seasons throughout 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, i.e., the exponential smoothing [33]. It is possible to define ω as follows:

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

where γ is the number of historical periods that are 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 $\lambda(t)_k$, it is possible to 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 is the number of time periods contained in a week.

C. ARIMA 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 instantly adapt to these changes.

The ARIMA [16] is a well-known methodology for both modeling and forecasting univariate time-series data such as traffic-flow data [26], electricity price [29], and other short-term prediction problems such as the one presented here. There are two main advantages to using ARIMA compared with other algorithms. First, it is versatile to represent very different types of time series, i.e., the autoregressive (AR) ones, the moving average (MA) ones, and a combination of those two (ARMA). Second, 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 nonseasonal stationary time series) is presented next, following the existing description in [30] (however, our framework can also detect both seasonal and nonstationary series). For a more detailed discussion, the reader should consult a comprehensive time-series forecasting text such as the one presented in [31].

In an ARIMA model, the future value of a variable is assumed to be a linear function of several past observations and random errors. It is possible to formulate the underlying process that generates the time series (taxi service over time for a given stand k) as

$$R_{k,t} = \kappa_0 + \phi_1 X_{k,t-1} + \phi_2 X_{k,t-2} + \cdots + \phi_p X_{k,t-p} + \varepsilon_{k,t} - \kappa_1 \varepsilon_{k,t-1} - \kappa_2 \varepsilon_{k,t-2} - \cdots - \kappa_q \varepsilon_{k,t-q} \tag{7}$$

where $R_{k,t}$ and $\{\varepsilon_{k,t}, \varepsilon_{k,t-1}, \varepsilon_{k,t-2}, \ldots\}$ are the actual value at time period t and the *Gaussian white noise* error terms observed in the past signal, respectively; $\phi_l (l = 1, 2, \ldots, p)$ and

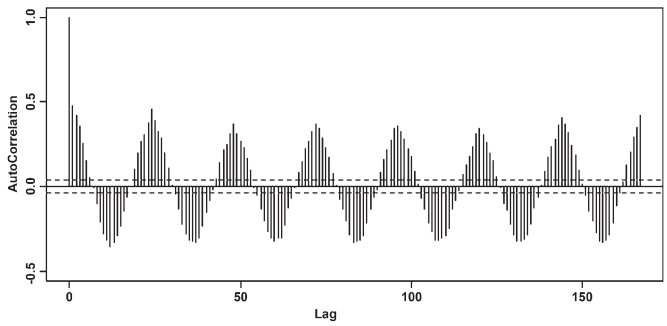

Fig. 2. Autocorrelation profile for data on the demand for taxi service that is obtained from one of the busiest taxi stands in the city (periods of 60 min). The x-axis has different period lags studied, and the y-axis contains the correlation within the signal. Note that there are peaks for each 12-h period.

$\kappa_m (m = 0, 1, 2, \ldots, q)$ are the model parameters/weights; and p and q are positive integers that are 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 proposed by Box *et al.* in [16]. They are useful for detecting the signal is periodic and, most important, which are the frequencies of these periodicities. A study conducted on time series from the demand of taxi services in one of the busiest taxi stands is presented in Fig. 2.

D. Sliding-Window Ensemble Framework

Three distinct predictive models have been proposed, which focus on learning from the long-, mid-, 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 community attention due to their drifting characteristics. The ensembles of such models were specifically addressed due to the challenge that is related to this type of data. One of the most popular models is the weighted ensemble [34]. The model proposed next is based on this one.

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

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

where ρ_{iH} is the error of 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) comparatively to the real service count time series. As the information is continuously arriving for the next periods $t, t + 1, t + 2, \ldots$, the window will also **slide** to determine how the models are performing in the **last** H **periods**.

To calculate such error, the symmetric mean percentage error (sMAPE) was used, which is formally described in Section V.

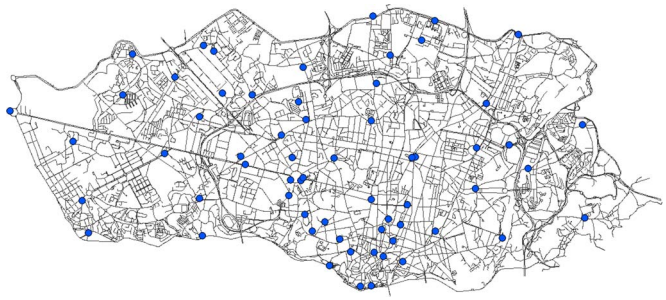

Fig. 3. Taxi-stand spatial distribution in the city of Porto, Portugal.

IV. Data Acquisition and Preprocessing

The stream event data of a taxi company operating in the city of Porto, Portugal, was used as the case study. This city is at 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. According to a recent aerial survey of the road traffic of the city [35], taxis represent 4% of the running vehicles during a nonrush-hour period. The existing regulations force the drivers not to *randomly* run in search for passengers; instead, they must choose a specific taxi stand out of the 63 existing ones in the city and to wait for the next service immediately after the last passenger drop-off. A map of the stands' spatial distribution is presented in Fig. 3.

The three main ways to pick up a passenger are as follows: 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 requests a taxi for a specific location/time; the parked taxis have priority over the running vacant ones in the central taxi dispatch system; and 3) a passenger picks a vacant taxi while it is going to a taxi stand on any street.

This section describes the company that is studied, the data acquisition process, and the preprocessing method that is applied.

A. Data Acquisition

The data were 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 8-h shifts, i.e., from midnight to 8 AM, from 8 AM to 4 PM, and from 4 PM to midnight. Each data chunk arrives with the following six attributes: 1) TYPE, which is relative to the type of event reported (it has four possible values: *busy*, i.e., the driver picked up a passenger; *assign*, i.e., the dispatch central assigned a previously required service; *free*, i.e., the driver dropped off a passenger; and *park*, i.e., the driver parked at a taxi stand); 2) STOP, which is an integer with the ID of the related taxi stand; 3) TIMESTAMP, which is the date/time in seconds of the event; 4) TAXI, which is the driver code; and 5) LATITUDE and 6) LONGITUDE, corresponding to the acquired GPS position. The data were acquired over a nonstop period of nine months. This paper only uses as input/output

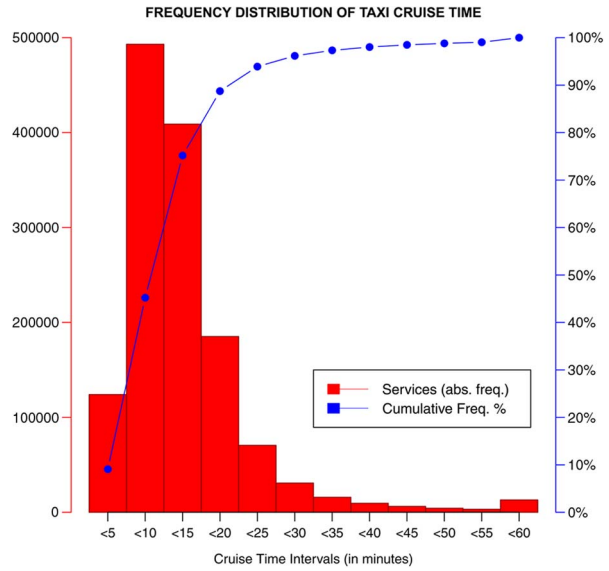

Fig. 4. Frequency distribution of taxi cruise time.

the services that are directly obtained at the stands or those automatically dispatched to the parked vehicles (more details in the succeeding section). This was done because the passenger demand at each taxi stand is the main feature for aiding the taxi drivers' decisions.

B. Preprocessing and Data Analysis

A time series of taxi demand services that are aggregated for a period of P minutes was developed as preprocessing. The three types of events are the following: 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. Both type-1 and type-2 events were considered service required. However, for each type-2 event, the system receives a *busy* event a few minutes later, i.e., as soon as the driver effectively picks up the passenger, which 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 a type-1 event related to the nearest taxi stand according to the defined criteria. This was done because many regulations prohibit passengers from being picked up in a predefined radius around a stop (in Porto, a 50-m radius is in place). Statistics about the period studied are presented. Fig. 4 presents the sample distribution of the cruise time of the services required. Table I details the number of taxi services demanded per daily shift and day type. Table II contains information about all services per taxi/driver and cruise time. The *service* column in Table II 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 directly requested on the street), while 77% of the service is directly demanded to taxis that are parked in a taxi stand (and 23% is assigned while they are cruising). The average waiting time (to pick up passengers) of a taxi that is parked at a taxi stand is

TABLE I
TAXI SERVICES VOLUME (PER DAY TYPE/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

42 min, while the average time for a service is only 11 min and 12 s. 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 that is presented here, where the shortness of services could be mitigated by obtaining services from the competitors.

The data in Tables I and II sustain that, despite the regularity in the service (particularly on weekends), there are major differences among the services that are provided by each driver (i.e., a large variance in service number and profit) related to their different levels of mobility intelligence. Fig. 4 focuses on the length of the services; 75% of them last 15 min or less. These statistics sustain the importance of a smart decision on the stand-choice problem; an accurate sensor to measure the passenger demand can be a major advantage in urban areas where a highly competitive scenario, like the one described here, is in place.

V. EXPERIMENTAL RESULTS

This section first describes the experimental setup that is developed to test the model on the available data. Second, the metrics that are used to evaluate the methods are enumerated. Finally, the results that are achieved are presented and discussed.

A. Experimental Setup

The testbed was based on the *prequential* evaluation [36]; data on the events that were occurring in the network was continuously acquired. An H-sized sliding window was used to measure the model error before each new prediction about the service count for the next period (the metrics that are used to do so are defined in Section IV-B). Each new real count was used to update the forecasting 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 central processing unit (CPU; the time series for each stand is independent from the remaining ones), which reduced the computational time required for each forecast. Fig. 5 shows the testbed that is described;

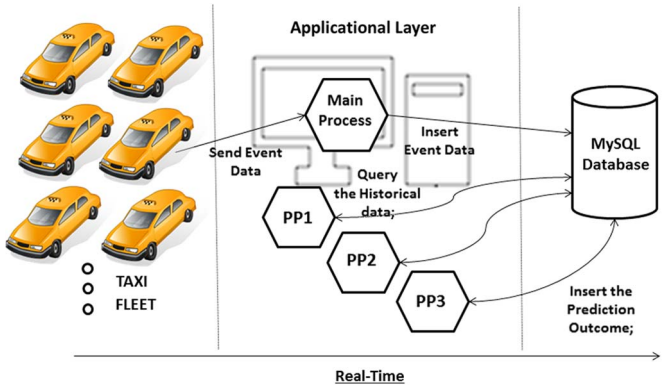

Fig. 5. Illustration of the streaming testbed.

$PP_i \ldots PP_t(t = 3)$ are the independent predicting processes, and each one handles a predetermined group of taxi stands. The predefined functions that are used and the values that are set for the model parameters are described in detail along this section.

An aggregation period of 30 min was set (i.e., a new forecast is produced every 30 min; $P = 30$) and a radius of 100 m ($W = 100 > 50$ 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 min.

The ARIMA model (p, d, and q values, and seasonality) was first set (and updated each 24 h) by learning/detecting the underlying model (i.e., autocorrelation and partial autocorrelation 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, an automatic time-series function in the [forecast] R package [38], i.e., *auto-arima*, was employed 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 nonweighted) were also updated every 24 h. A sliding window of 4 h ($H = 8$) 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). The goal was to calibrate the model by finding the optimal subregion on the input space $\alpha \in [0, 1]$, which maximizes the predictive performance. To do so, 100 distinct samples were generated as admissible values for α, and they were tested using an older and smaller data set containing data very similar to the one tested in our experiments (i.e., the same feature space). As a result, it was possible to determine the ideal value as $\alpha = 0.4$. This value demonstrated to be robust, i.e., changes did not cause a significant impact on the model output since they remain stable on the input space of 0.4 ± 0.1. Therefore, $\alpha = 0.4$ was used in the 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 weight $\omega_{i>\gamma} \sim 0$. According to this, $\alpha = 0.4 \Longrightarrow \gamma = 8$.

Table III summarizes the information about the learning periods that are 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

The data that were obtained from the last four months were used to evaluate the framework (where 506 873 services emerged). A well-known error measurement was employed to evaluate the output, i.e., the sMAPE [40]. This metric is formally defined as follows.

Consider $R = \{R_{k,1}, R_{k,2}, \ldots, R_{k,t}\}$ 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 period $[1, t]$ and $X = \{X_{k,1}, X_{k,2}, \ldots, X_{k,t}\}$ to be the number of services that actually emerged in the same conditions. sMAPE_k, i.e., the error that is measured on the time series of services that are predicted to stand k, can be defined as

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

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

where t is the number of time periods that are considered. However, this metric can be too intolerant to small magnitude errors (e.g., if two services are predicted on a given period for a taxi stand of interest but no one emerges, the error that is measured during that period would be 1). To produce more accurate statistics about the series containing very small numbers, this was performed by adding a Laplace estimator [41] to (10). In this case, we will do it by adding constant c to the denominator (i.e., originally, it was added to the numerator to estimate a success rate [41]). Therefore, it is possible to redefine sMAPE_k as follows:

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

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

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

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

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

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%**

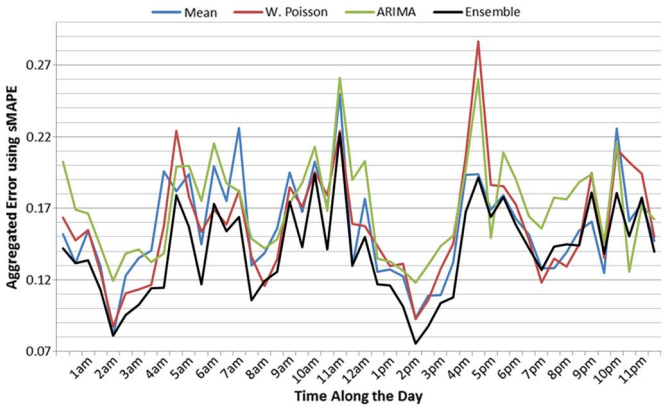

Fig. 6. Ensemble evaluation on a typical Saturday.

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

C. Results

The results are presented over the following 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 output examples; and 4) a report on the computational time that is required to forecast the next period.

First, the error that is measured for each model is shown in Table IV. The results are first presented per shift and then globally. The results were aggregated using AG_β which was previously defined.

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

Third, three distinct weekly analyses of the discrepancies between the demand that is predicted and the services that are actually provided are shown in Fig. 7. The model forecasted the spatiotemporal taxi–passenger demand for every 30-min period using (on average) 38.12 s of processing time (i.e., 1.906 seconds per time series/stand) as a result of the computational parallel approach that was presented before. This method reduced the computational time by 70% (i.e., in the first three weeks, the model was tested using just one iterative process—one program and one CPU core—and on average, the computational time was 99.77 s). The ARIMA model update was also fast, i.e., 48.12 s (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 that is considered; the error measured was always lower than 26%. The models just present slight discrepancies within the daily shifts.

The ensemble methodology is comparatively robust to the remaining models; in Fig. 6, it is possible to identify a point where the ensemble maintained its performance while two other methods suffered a significant decrease in performance, highlighting the inherited learning of the ensemble approach. Fig. 7 shows two distinct scenarios to compare the forecasted and real demands. In Fig. 7(a), the demand corresponds to an irregular taxi stand where services do not have a usual pattern to emerge (even if the demand is low); in Fig. 7(b), the chart corresponds to a completely regular stand behavior. The two examples illustrate that the ensemble can correctly forecast the demand in distinct scenarios, periods, and time horizons. In the scenario that is presented, the target variable is the number of services to arise at a taxi-stand network during a predefined period of time. This variable was chosen due to the **stand** relevance in this scenario (where 76% of the total number of services is directly required to vehicles that are parked on them). However, this is not the reality in many large cities around the world due to their (de)regulation [6]. Most authors in the literature on this topic divide their scenarios/urban areas into spatial clusters, as shown 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]. The mathematical model does not depend on how the service historical data are 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.

VI. CONCLUSION AND FUTURE WORK

This paper has presented a **novel application of time-series forecasting techniques to improve taxi-driver mobility intelligence**. Thjs was done by transforming both the 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 a time series of interest to use first as a learning base for our model and second as a streaming test framework. As a result, the model that is presented has been able to predict the taxi–passenger demand at each one of the 63 taxi stands for 30-min period intervals.

The model has presented a more than satisfactory performance, correctly predicting the 506 873 tested services with an aggregated error measurement lower than 26%. It is our belief that **this model is a true novelty** and **a major contribution** to the area due to its adapting characteristics:

1) It mines both the periodicity and the seasonality of the passenger demand, regularly updating itself.
2) It simultaneously uses long-, mid-, and short-term historical data as a learning base.

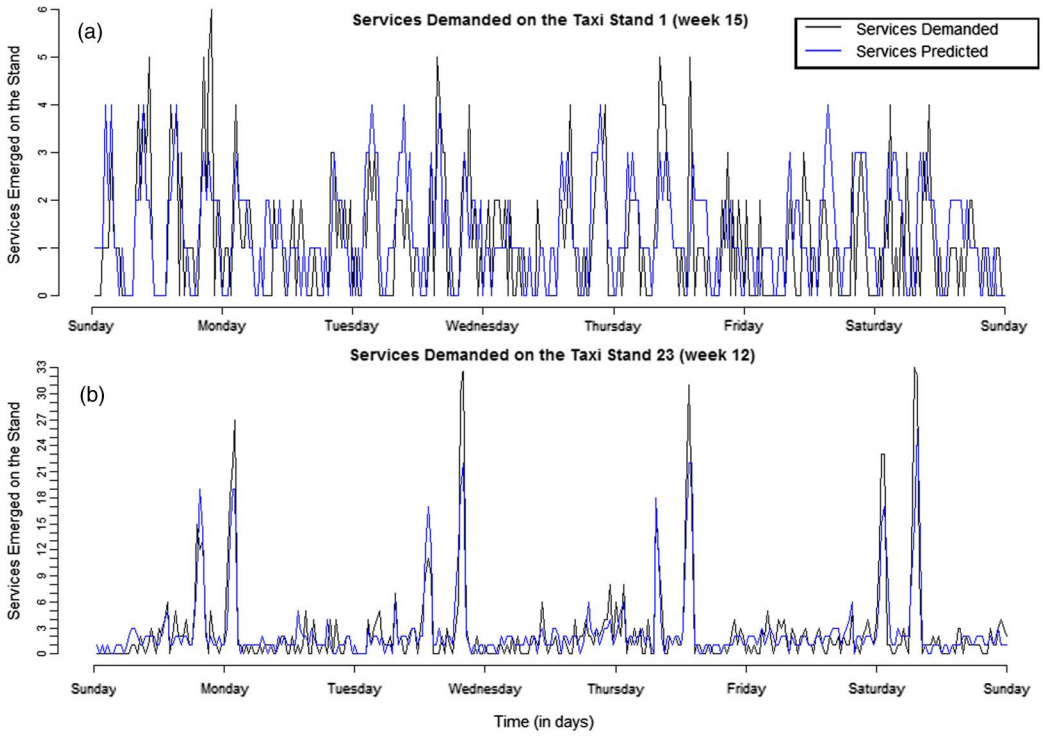

Fig. 7. Weekly comparison between the services that are forecasted and the services that emerged on two distinct scenarios/taxi stands and weeks.

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

3) It takes advantage of the ubiquitous characteristics of a taxi network, assembling the experience and the knowledge of all vehicles/drivers, which they usually just do on their own.

This approach meets no parallel in the literature due to its testbed; the models have been tested in a streaming environment, while the state of the art mainly presents offline experimental setups.

This model will be used as a feature for a recommendation system (to be developed), which will produce smart live recommendations to the taxi drivers about which taxi stand they should head to after a drop-off. This decision support framework will also address other features such as distance or live traffic conditions. We believe that the deployment of such a system in a taxi fleet will contribute to increasing its competitiveness facing other taxi fleets in a Scenario-1 network (e.g., such as the network that was studied, 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.

ACKNOWLEDGMENT

The authors would like to thank Geolink, Lda. and its team for the data they supplied, and the anonymous reviewers for their valuable comments and suggestions for improving this paper.

REFERENCES

[1] A. Glaschenko, A. Ivaschenko, G. Rzevski, and P. Skobelev, "Multi-agent real time scheduling system for taxi companies," in *Proc. 8th Int. Conf. AAMAS*, Budapest, Hungary, 2009, pp. 29–36.

[2] J. Lee, G.-L. Park, H. Kim, Y.-K. Yang, P. Kim, and S.-W. Kim, *A Telematics Service System Based on the Linux Cluster*. Berlin, Germany: Springer-Verlag, 2007, pp. 660–667.

[3] J. Yuan, Y. Zheng, X. Xie, and G. Sun, "T-drive: Enhancing driving directions with taxi drivers' intelligence," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 1, pp. 220–232, Jan. 2013.

[4] P.-Y. Chen, J.-W. Liu, and W.-T. Chen, "A fuel-saving and pollution-reducing dynamic taxi-sharing protocol in VANETs," in *Proc. 72nd IEEE VTC-Fall*, 2010, pp. 1–5.

[5] P. d'Orey, R. Fernandes, and M. Ferreira, "Empirical evaluation of a dynamic and distributed taxi-sharing system," in *Proc. 15th IEEE Int. Conf. ITSC*, Sep. 2012, pp. 140–146.

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

[7] H. Yang, K. I. Wong, and S. C. Wong, "Modeling urban taxi services in road networks: Progress, problem and prospect," *J. Adv. Transp.*, vol. 35, no. 3, pp. 237–258, Fall 2001.

[8] Z. Deng and M. Ji, "Spatiotemporal structure of taxi services in Shanghai: Using exploratory spatial data analysis," in *Proc. 19th Int. Conf. Geoinf.*, 2011, pp. 1–5.

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

[10] Y. Yue, Y. Zhuang, Q. Li, and Q. Mao, "Mining time-dependent attractive areas and movement patterns from taxi trajectory data," in *Proc. 17th Int. Conf. Geoinf.*, 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 *Proc. IEEE Int. Conf. PERCOM Workshops*, Mar. 2011, pp. 63–68.
[12] J. Lee, I. Shin, and G. Park, "Analysis of the passenger pick-up pattern for taxi location recommendation," in *Proc. 4th Int. Conf. NCM Adv. Inf. Manage.*, 2008, vol. 1, 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," in *Proc. Ambient Intell.*, 2010, vol. 6439, pp. 86–95.
[14] J. Gama, *Knowledge Discovery From Data Streams*. London, U.K.: Chapman & Hall, 2010.
[15] A. Ihler, J. Hutchins, and P. Smyth, "Adaptive event detection with time-varying Poisson processes," in *Proc. 12th Int. Conf. ACM SIGKDD*, 2006, pp. 207–216.
[16] G. Box, G. Jenkins, and G. Reinsel, *Time Series Analysis*. San Francisco, CA, USA: Holden-Day, 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," *Front. Comput. Sci. Chin.*, vol. 6, no. 1, pp. 111–121, Feb. 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," *Perv. Mobile Comput.*, vol. 6, no. 4, pp. 455–466, Aug. 2010.
[19] H. Chang, Y. Tai, and J. Hsu, "Context-aware taxi demand hotspots prediction," *Int. J. Business Intell. Data Mining*, vol. 5, no. 1, pp. 3–18, Dec. 2010.
[20] B. Cule, B. Goethals, S. Tassenoy, and S. Verboven, "Mining train delays," in *Advances in Intelligent Data Analysis X*. Berlin, Germany: Springer-Verlag, 2011, ser. LNCS, 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 *Proc 13th IEEE Conf. 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*. New York, NY, USA: Springer-Verlag, 2012, ser. LNCS, 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, Jun. 2008.
[24] J. Gama and P. Rodrigues, "Stream-based electricity load forecast," in *Knowledge Discovery in Databases: PKDD*. Berlin, Germany: Springer-Verlag, 2007, ser. LNCS, 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," *J. Transp. Eng.*, vol. 129, no. 6, pp. 664–672, Nov. 2003.
[26] W. Min and L. Wynter, "Real-time road traffic prediction with spatio-temporal correlations," *Transp. Res. C, Emerg. Technol.*, vol. 19, no. 4, pp. 606–616, Aug. 2011.
[27] Y. Ge, H. Xiong, A. Tuzhilin, K. Xiao, M. Gruteser, and M. Pazzani, "An energy-efficient mobile recommender system," in *Proc. 16th ACM SIGKDD Int. Conf.*, 2010, pp. 899–908.
[28] J. Yuan, Y. Zheng, L. Zhang, X. Xie, and G. Sun, "Where to find my next passenger?" in *Proc. 13th ACM Int. Conf. UbiComp*, 2011, pp. 109–118.
[29] J. Contreras, R. Espinola, F. J. Nogales, and A. J. Conejo, "ARIMA models to predict next-day electricity prices," *IEEE Trans. Power Syst.*, vol. 18, no. 3, pp. 1014–1020, Aug. 2003.
[30] G. Zhang, "Time series forecasting using a hybrid ARIMA and neural network model," *Neurocomputing*, vol. 50, pp. 159–175, Jan. 2003.
[31] J. Cryer and K. Chan, *Time Series Analysis With Applications in R*. New York, NY, USA: Springer-Verlag, 2008.
[32] L. Moreira-Matias, J. Gama, M. Ferreira, and L. Damas, "A predictive model for the passenger demand on a taxi network," in *Proc. 15th IEEE Int. Conf. ITSC*, Sep. 2012, pp. 1014–1019.
[33] C. Holt, "Forecasting seasonals and trends by exponentially weighted moving averages," *Int. J. Forecast.*, vol. 20, no. 1, pp. 5–10, Jan.–Mar. 2004.
[34] H. Wang, W. Fan, P. Yu, and J. Han, "Mining concept-drifting data streams using ensemble classifiers," in *Proc. 9th ACM SIGKDD Int. Conf.*, 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 *Proc. 6th ACM Int. Workshop Veh. InterNetw.*, 2009, pp. 53–62.
[36] A. Dawid, "Present position and potential developments: Some personal views: Statistical theory: The prequential approach," *J. Roy. Stat. Soc. A*, vol. 47, no. 2, pp. 278–292, Jan. 1984.
[37] R Core Team. (2012). R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria. [Online]. Available: http://www.R-project.org
[38] K. Yeasmin and J. H. Rob (2008, Jul.). Automatic Time Series Forecasting: The forecast package for R. *J. Stat. Softw.* [Online]. *27(3)*, pp. 1–22. 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," *Proc. IEEE*, vol. 95, no. 5, pp. 899–924, May 2007.
[40] S. Makridakis and M. Hibon, "The M3-competition: Results, conclusions and implications," *Int. J. Forecast.*, vol. 16, no. 4, pp. 451–476, Jan. 2000.
[41] E. Jaynes, *Probability Theory: The Logic of Science*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

Luis Moreira-Matias received the M.Sc. degree in informatics engineering from the University of Porto, Porto, Portugal, in 2009, where he is currently working toward the Ph.D. degree in machine learning with the Faculty of Engineering.

He is also with the Laboratory for Artificial Intelligence and Decision Support, Tecnologia e Ciência, Instituto de Engenharia de Sistemas e Computadores, University of Porto. His current research interest is learning from data streams.

João Gama received the Ph.D. degree in computer science from the University of Porto, Porto, Portugal, in 2000.

He is a Researcher with the Laboratory of Artificial Intelligence and Decision Support, Tecnologia e Ciência, Instituto de Engenharia de Sistemas e Computadores, University of Porto. He has recently authored a book on knowledge discovery from data streams. His main research interest is learning from data streams.

Michel Ferreira received the Ms.c. and Ph.D. degrees in computer science from the University of Porto, Porto, Portugal, in 1994 and 2002, respectively.

He is an Assistant Professor with the Faculty of Sciences, University of Porto, where he leads the Geo-Networks group with the Department of Computer Science. He has led several research projects in the areas of logic-based spatial databases, vehicular sensing, and intervehicle communication.

João Mendes-Moreira received the Ph.D. degree in engineering sciences from the University of Porto, Porto, Portugal.

He is an Assistant Professor with the Department of Informatics Engineering, Faculty of Sciences, University of Porto, where he is also a Researcher with the Laboratory for Artificial Intelligence and Decision Support, Tecnologia e Ciência, Instituto de Engenharia de Sistemas e Computadores. His research works focus on applied machine learning.

Luis Damas received the Ph.D. degree in mathematical theory of computation from the University of Edinburgh, Edinburgh, U.K., in 1984.

He is the creator of one of the most widely used logic programming systems, i.e., the YAP Prolog compiler. He is also the Co-founder and the Chief Technical Officer with Geolink, Lda., Porto, Portugal. His research interests include simulation, distributed systems, and programming languages.