2014 IEEE 17th International Conference on
Intelligent Transportation Systems (ITSC)
October 8-11, 2014. Qingdao, China

An Online Learning Framework for Predicting the Taxi Stand's Profitability

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

Abstract— **Taxi services play a central role in the mobility dynamics of major urban areas. Advanced communication devices such as GPS (Global Positioning System) and GSM (Global System for Mobile Communications) made it possible to monitor the drivers' activities in real-time. This paper presents an online learning approach to predict profitability in taxi stands. This approach consists of classifying each stand based according to the type of services that are being requested (for instance, short and long trips). This classification is achieved by maintaining a time-evolving histogram to approximate local probability density functions (p.d.f.) in service revenues. The future values of this histogram are estimated using time series analysis methods assuming that a non-homogeneous Poisson process is in place. Finally, the method's outputs were combined using a voting ensemble scheme based on a sliding window of historical data. Experimental tests were conducted using online data transmitted by 441 vehicles of a fleet running in the city of Porto, Portugal. The results demonstrated that the proposed framework can provide an effective insight on the characterization of taxi stand profitability.**

I. Introduction

Today, taxi services play a central role in the mobility dynamics of major urban areas. It offers a personalized destination service in a fast and yet secure manner. One of the largest companies operating in New York City performed roughly 470,000 trips in 2006, generating $1.82 billion US dollars in revenue. It represented 30% of the total public transportation fares and an averaged driver income per shift of $158 US dollars [1]. However, this level of revenue was achieved by performing an inefficient use of the resources available, namely the vehicles and fuel. Traditionally, taxi drivers earn their profit by *randomly* cruising the road network looking for a passenger. This naive strategy leads to large fuel wastes on heavily congested traffic and, consequently, to a low ratio of live miles (miles with a fare) over cruising miles (miles without a fare).

Experienced drivers are able to make smarter decisions regarding the taxi-passenger finding problem by *knowing* some demand patterns in advance. However, most drivers choose to use only a few number of stands to wait for their next service. Advanced communication devices such as the GPS (Global Positioning System) and the GSM (Global System for Mobile Communications) made it possible to **monitor** the drivers' activities in real-time. The data acquired by these systems can be used to combine the drivers' experience with intelligent decision support frameworks in order to recommend the most suitable area/stand to go after a passenger drop-off. This problem is known as the **taxi stand choice problem** [2]–[6]. The real-time stand-choice problem is based on four key variables: the expected revenue for a service over time, the distance/cost relation of each stand, the number of taxis already waiting at each stand and the passenger demand for each stand over time.

Most of the existing research on this topic focus on predicting passenger demand when characterizing stand profitability [4]–[6]. In fact, this can be true to a certain extent if the expected demand is the main variable in this problem. Knowing where the demand will occur is a valuable contribution to reducing cruising miles. However, a question arises: is this knowledge **enough**? There are roughly two types of taxi networks: the ones where service demand is larger than the supply (Scenario 1), and the ones where the opposite happens (Scenario 2). In the first scenario, the answer to that may be positive. However, that does not happen in the second scenario where the drivers have to **select** which services to take and which services to ignore.

This paper presents an online learning approach to predict taxi stand profitability. This approach consists of classifying each stand based on the type of services demanded (for instance, short or long trips). By doing so, the authors expect to provide a framework capable of giving a continuous short-term perspective on which are the stands where high-profit services will be demanded. This stepwise approach starts by maintaining a time-evolving histogram to approximate local probability density functions (p.d.f.) in service revenues. The p.d.f. curves are used to classify each stand over time. Then, the short-term future values of the histogram are estimated using time series analysis approaches based on non-homogeneous Poisson processes [4]–[6]. Finally, the outputs of the methods were combined using a simple voting ensemble scheme.

Manuscript submitted April 10, 2014 and revised on August 18, 2014.

This work was supported by projects VTL - "Virtual Traffic Lights", MAESTRA and I-CITY - "ICT for Future Mobility", as part of Grants PTDC/EIACCO/118114/2010, ICT-2013-612944 and NORTE-07-0124-FEDER-000064, respectively. The authors also acknowledge the support of the ERDF - European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness), and of Portuguese Funds through the FCT (Portuguese Foundation for Science and Technology) as part of project FCOMP-01-0124-FEDER-037281.

Luis Moreira-Matias is with Faculdade de Engenharia, U.Porto 4200-465 Porto, Portugal and with Instituto de Telecomunicações , 4200-465 Porto, Portugal (phone: 00351-91-4221647; e-mail: luis.matias[at]fe.up.pt).

João Gama is with LIAAD-INESC TEC and U.Porto, 4200-465 Porto - Portugal (e-mail: jgama[at]fep.up.pt).

Michel Ferreira is with the Instituto de Telecomunicações, U.Porto, 4169-007 Porto - Portugal (e-mail: michel[at]dcc.fc.up.pt).

João Mendes-Moreira is with LIAAD-INESC TEC and with U.Porto 4200-465 Porto - Portugal (e-mail: jmoreira[at]fe.up.pt).

Luis Damas is with Geolink, Lda., Avenida de França, 20, Sala 605, 4050-275 Porto – Portugal (e-mail: luis[at]geolink.pt).

978-1-4799-6077-4/14/$31.00 ©2014 IEEE

A large 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 running one fleet. The data transmitted between August 2011 and February 2012 by the largest company, which has 441 vehicles, was used as test bed for the methodology presented here. In this scenario, the average cruising time in each service is $\sim$ 12 minutes. Consequently, 65% of the services represent low revenues ($<$ \$8 US dollars). The voting-based ensemble outperformed the remaining predictive methods used on the testbed scenario by achieving an average accuracy of 74%. These results support the concept hereby presented where each stand profitability is given by the **size** of their short-term service revenues.

The remainder of the paper is structured as follows: Section 2 revises the existing literature on this topic. Section 3 formally presents the model employed. Section 4 firstly describes how the dataset used was acquired and preprocessed. Then, some statistics about the dataset are presented. Section 5 describes how the methodology was tested in a real scenario: firstly, the experimental setup and metrics used to evaluate the model are introduced; then, the results obtained are presented in detail, followed by some important remarks. Finally, in the last section conclusions are drawn and topics for future work are outlined.

II. Related Work

More and more datasets containing historical GPS data sets are being explored to improve taxi driver profitability. Typically, studies employ one of the following approaches: (1) predicting the number of service requests within a given area or (2) selecting some areas where there will be a high demand for services in the short-term. In (1), the state-of-the-art approaches are time series analysis techniques, namely the Poisson-based Autoregressive Integrated Moving Averages (ARIMA) [4]–[6] and Exponential Smoothing [5], [6]. In [4], an extended ARIMA model is used where a time-varying Poisson process is assumed to be in place. The work by Moreira-Matias *et al.* [5] extends this concept by introducing both an online ensemble scheme and a Poisson-based Exponential Smoothing, which uses historical data to build the predictive model. The same authors extended the methodology to be fully incremental [6] by proposing a perceptron-based rule to update the ARIMA weights for each prediction instead of finding its optimal fitting.

Type-2 approaches rely on recommending highly profitable routes. The main goal of these routing techniques is to establish Origin-Destination matrices to select demand hotspots (regions that are more likely to provide high demand rates). Hierarchical clustering is employed in [7], [8] while the work in [8] also explores DBSCAN to mine time-dependent attractive areas by analyzing the historical time series of demands within predefined time spans. These approaches were extended by Hu *et al.* [9] who proposed a heuristic function to connect the centroids of the top-k hotspots and a probability model to estimate the gasoline consumption in every route to compute the link weights.

The abovementioned approaches aim at increasing the ratio between live and cruising miles. However, this may be misleading as the variability in service revenue is high, especially from region to region [10]. Let us formulate this issue with a numerical example: a predictive model of interest forecasted a demand of $d_1 = 10$ and $d_2 = 6$ services in areas/stands A_1, A_2, respectively, during the following period of P minutes. Let C_1, C_2 denote the number of cars already parked in the stands. The profit at each stand can be expressed as follows:

$$A_{profit} = r - \left(\frac{C \times P}{d} \times \tau\right) \quad (1)$$

where r is the expected service revenue and τ is a **constant** expressing the cost of letting a vehicle wait in line at a stand per unit of time. Assuming that both are equally distant from our current location and that the number of vehicles already parked in those areas is similar (i.e. $C_1 \sim C_2$), it is possible to estimate that the relationship between waiting times to pick-up the next passenger at each stand δ_1, δ_2 will be $\delta_1 = 0.6 \times \delta_2$. Finally, if the waiting time cost is considered to be independent from the area under analysis and that the average revenue at each stand r_1, r_2 is, for instance, \$8 and \$14, the most profitable stand would be A_2 and not A_1. A typical example of this could be airports, where long-runs are normally provided from city outskirts to downtown areas.

The work in [4] presented a more accurate approach to the profitability problem by profiling the driver's experience according to the historical data on high-profit/low-profit drivers. However, the work that comes closest to the one presented here is the work by Powell *et al.* [10], who present a model to estimate the most profitable route by employing a spatial window to model the profitability of the neighboring regions, regarding the short-term decision on the path to take. The area's revenue scores end up being computed based on a moving average of the fares using a very short time window (i.e. 60 minutes). However, previous work has already shown how important the mid and long-term history can be to compute demand-based predictions [5], [6]. Moreover, by maintaining the fares as continuous variables, the authors oversimplify the concept of "low/high" fare to make it constant, which can be misleading (e.g. a \$10 dollars service may not be relevant on the morning peak but can be valuable on the evening one; a peak value can be harder to predict than a class label). By maintaining a fair approximation to the revenue p.d.f., the present approach should be **adaptable** to every scenario, allowing the user to decide which should be the rules in place to consider a service revenue *high*. Moreover, it combines sliding windows of different lengths to explore the historical data on different levels. For these reasons, the present framework meets no parallel in the existing literature on this topic.

III. Methodology

Let $X_k = \{X_{k,0}, X_{k,1}, ..., X_{k,t}\}$ be a discrete time series (aggregation period of P-minutes) for the number of services requested at a taxi stand k. Let $R_{k,t}$ denote a vector containing the revenue values corresponding to the amount paid by each service which starts at the stand k at time period t, where $X_{k,t} = |R_{k,t}|$. To characterize the distribution of these values, the authors propose to approximate its local p.d.f.. One of the best known ways of doing that is by discretizing the variable values into intervals using histograms [11]. By dividing the number of services $X_{k,t}$ into n bins according to service revenue, it is possible to obtain n discrete time series for the number of services requested within a certain *revenue interval*. Secondly, a set of fixed rules is employed to classify the period's profitability based on those histograms. Thirdly, time series analysis techniques are employed to estimate the future values of these n series based on previous work about demand prediction. Those values will be used to predict the stand's short-term profitability class by employing the abovementioned set of rules. Finally, an ensemble voting scheme is employed to combine each method's prediction into the final one. This methodology is illustrated in Fig. 1. Its details are provided along this section.

A. On Discretizing the Revenues

The first goal is to discretize the revenues into a value *interval* $\pi_i = [b_i, b_{i+1}) \in \Pi$ for $R_{k,t}$ such that $b_i \leq R_{k,t} < b_{i+1}$. Π can be defined as follows

$$\Pi = \{\pi_i | \pi_i = [b_i, b_{i+1}) : b_{i+1} - b_i = b_i - b_{i-1}, \forall b_i \in N\} \tag{2}$$

where $\delta = b_{i+1} - b_i$ denotes the interval **width**. Consequently, it is possible to obtain an *equal-width* histogram $h(F, B)$ defined by the aforementioned set of break points $B = b_1, ..., b_{n-1}$ and a set of frequency counts $F = f_1, ..., f_n$. To define the number of bins n, it is necessary to define the *range* of the random variable and the desired *interval width*. For that, three user-defined parameters are employed: the interval width μ and a minimum/maximum value as mi, ma, respectively. Therefore, it is possible to redefine π_i as follows:

$$\pi_i = [mi + \mu \times (i-1), mi + \mu \times i) : (mi + \mu \times i) \leq ma \tag{3}$$

An additional *last* bucket is added to the ones defined in Π to account for all the revenue values above the threshold value (i.e. ma). Consequently, $n = |\Pi| + 1$.

By employing these histograms, it is possible to monitor the evolution of the revenue's p.d.f. at a given taxi stand to predict the short-term one . Estimating the p.d.f. estimation brings a vast range of possibilities when it comes to building a set of rules (or multiple rules) capable of classifying the stand's profitability in every time period. The set of rules used in this particular scenario is described in Section V-A.

B. Numerical Predictions for the Stand's Classification

Regardless of the evolution of the p.d.f. throughout time, the number of bins n is constant over time (it only depends on the parameters ma, mi and μ). Consequently, each bin can be seen as a time series in terms of the number of services requested at that stand where the revenues are constrained by a given interval. This observation makes it possible to model the p.d.f. estimation problem as multiple time series forecasting ones . Departing from the previous work in [5], it is expected that these new series also follow non-homogeneous Poisson processes. Consequently, the three predictive models used in the previous work (ARIMA, Time-Varying Poisson Models and Exponential Smoothing) were also applied to these n time series using the very same parameter setting (namely, a two-week period to train both the ARIMA and the Exponential Smoothing). The result is a set of three p.d.f. which, based on the set of rule mentioned before, are capable of outputting three labels on the profitability of the next time period. A majority voting scheme was employed to combine the label outputs according to each prediction. This simple scheme consists of measuring the average accuracy of each method on the last δ periods. The average accuracy of each method is calculated every 24h. This ensemble is merely a wise selection of the best predictive method for each stand.

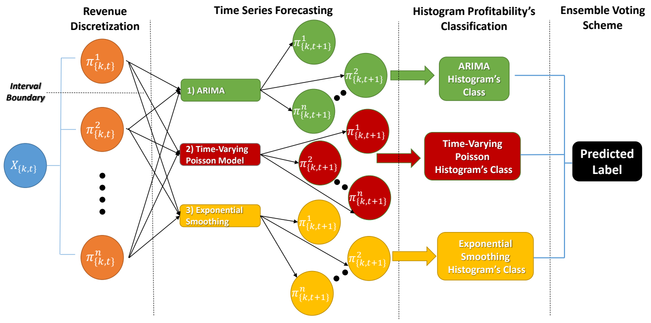

Fig. 1: Illustration of the predictive framework proposed.

However, since the algorithm's output is numerical, it is possible to explore other predictive approaches over the obtained p.d.f., which is commonly used in on machine learning research works. Nevertheless, the goal with this paper is to demonstrate its effectiveness in predicting the short-term stand profitability as a simple proof of concept.

IV. Case Study

A taxi company operating in the city of Porto, Portugal, was used as case study. This city is the center of a medium-sized urban area (consisting of 1.3 million inhabitants) where passenger demand is lower than the number of vacant taxis running, and this causes taxi companies to compete fiercely. The data were acquired using the telematics installed in each of the 441 running vehicles which are part of the company fleet. The data refer to a non-stop period of seven months between August 2011 and February 2012, containing nearly one million trip records. At first, each chunk of data comes with the following five attributes: the driver's ID, a Julian timestamp, the taxi status (zero/one for vacant/busy), and the latitude/longitude coordinates. By preprocessing the data, it was possible to obtain a dataset containing one entry per service where the variables were a 1) Julian timestamp marking the beginning of the trip, 2) the stand from which the taxi departed, 3) the cruising distance (in meters) and 4) the cruising time (in seconds). All services that did not start at a taxi stand were not considered in this study.

A simplified version of Porto's taxi service price structure was used to conduct these experiments, which is illustrated

TABLE I: Porto's taxi service price structure. Both the temporal and spatial fractions cost 0.15 euros.

Location	Time	Minimum Price	Minimum Distance	Spatial Fraction	Temporal Fraction
Inside the city limits	6am → 9pm	2.00eur.	220.0m	333.3m	37.0 sec.
	9pm → 6am	2.50eur.	176.0m	277.7m	37.0 sec.
Outside the city limits	6am → 9pm	3.25eur.	220.0m	166.6m	37.0 sec.
	9pm → 6am	3.90eur.	176.0m	138.9m	37.0 sec.

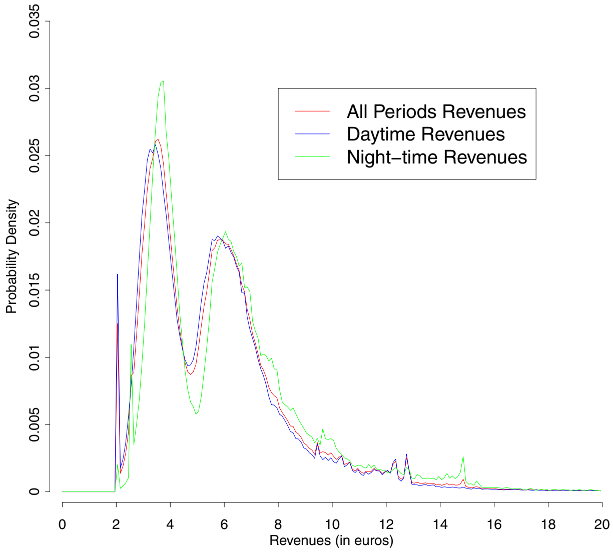

Fig. 2: Sample-based p.d.f. for the revenues detailed by daytime and nightime.

in Table I. Fig. 2 represents three sample-based estimations of the revenue's p.d.f.: a global estimation, one for the daytime revenues and another for the nighttime revenues. All estimations exhibit a **bimodal** structure. That is even clearer when the nightime scenario is analyzed. The time lag between the nighttime and the remaining p.d.f. indicates that the nighttime services usually have larger revenues than daytime services. Fig. 3 illustrates an equal-width revenue histogram and its cumulative frequency. Note that nearly 60% of the demanded services have a revenue below 6 euros. This pattern shows how difficult it is to maintain a balanced relationship between service offer and demand in this particular case study.

V. Experiments

A. Experimental Setup

The parameter setting is described in Table II. The first six months of data were used to train the model, while the last month (February) was used as test set.

The numerical prediction methods setup followed closely the one proposed in a previous study [5]. The ARIMA model (p, d, q values and seasonality) was firstly set (and updated each 24h) by learning/detecting the underlying model running on the historical time series curve of each stand during the last two weeks. For that, an automatic time series function in the [forecast] R package [12] - *auto-arima* was employed. 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 updated every 24h.

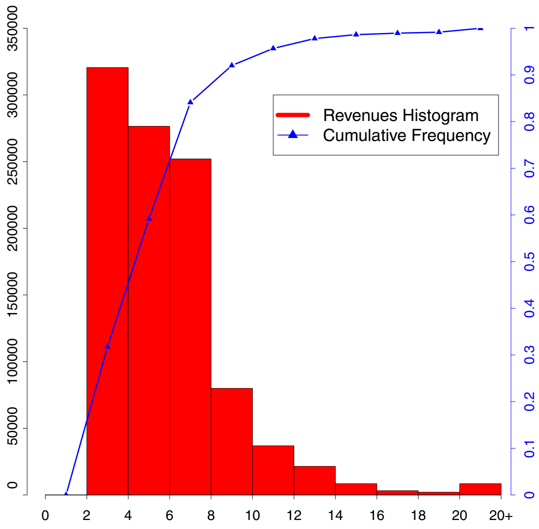

Fig. 3: Equal-Width Revenue Histogram and its cumulative frequency.

TABLE II: Parameter Setting used in the experiments.

Parameter	Value	Description
γ	8	Learning period for the Exponential Smoothing (8 weeks)
α	0.4	parameter to calculate the weight's curve on the Exponential Smoothing
P	60	aggregation period used to calculate the time series (in minutes)
mi	2	minimum value in the revenue histogram obtained for each period
ma	10	maximum bounded value in the revenue histogram obtained for each period
μ	4	bounded width of the intervals

The parameter setting for (ma,mi,μ) resulted in histograms with three bins (i.e. $n = 3$). For this particular scenario, the authors have established a three-class set to estimate the stands' profitability (”*low*”,”*medium*” and ”*high*”). A user-defined set of rules was developed for this particular task adapted to the present scenario. Its pseudo code is displayed in Fig. 4. However, the authors sustain that this approach can be adapted by any taxi network by changing n, the number of classes and the rule set in place.

The labels obtained with the three numerical predictions were compared to the ensemble method. A majority class baseline was also considered in the experimental setup. These five methods were tested on all 63 stands available.

B. Evaluation Metrics

The Root Mean Square Error (RMSE) and the Accuracy (ACC) were used as evaluation metrics. Moreover, the accuracy error was divided into `higher-prediction` and `lower-prediction` to discover when the predicted profitability is higher/lower than the real one.

These metrics were calculated for the N periods considered in the test set. They were then aggregated by calculating a *weighted* mean of their values at the existing taxi stands. Each stand's weight corresponds to the number of services requested on them.

```
function CLASSIFY-PERIOD($h(F, B), X_{k+t}$)
    if $X_{k+t} = 0$ then return "low";
    end if
    if $X_{k+t} <= 5$ then
        if $b_1 = 0$ then return "medium";
        else
            return "low";
        end if
    end if
    $b1_{ratio} = b_1/X_{k+t}$;
    if $b1_{ratio} < (1 - 0.4)$ then return "high";
    else
        return $b1_{ratio} < (1 - 0.2)$ return "medium";
    end if
    return "low";
end function
```

Fig. 4: Period Profitability Classification using the Revenue Histogram. The parameters represent the histogram's frequencies (F) and break points (B), as well as its total mass X_{k+t}.

C. Results

Fig. 5 presents descriptive statistics on the bin values of one of the busiest taxi stands in this case study. This statistics are divided by profitability class and also by day period. This division shows how the classification rule set (Fig. 4) works over the histograms. Using these rules, the following class distribution was obtained: *"low"*: 81.57%; *"medium"*: 13.10%; *"high"*: 5.33%. Table III presents a detailed evaluation of the five classification frameworks employed in this task. Finally, Fig. 6 divides the ensemble accuracy between each of the 63 taxi stands in Porto grouped with the number of services requested at the stand during the test period.

D. Discussion

The results in Table III demonstrate that the Ensemble method presents the highest accuracy in the time-dependent profitability classification task. It is important to highlight that this method surpasses the majority class method, especially if we consider that we are facing an unbalanced classification task (i.e. 81.57% of the true labels are *"low"*. Fig. 5 exemplifies the histograms distribution on distinct classes and scenarios. Note that the class ("low"/"medium"/"high") does not have a direct relationship with the bins frequencies.

To approximate p.d.f. using histograms may seem quite simpler while compared with other estimation methods (e.g. kernel estimation). It may partially explain the accuracy errors on the stand revenue's classification. However, this method have a strong advantage facing the most common ones: it can be computed nearly **incrementally**, using one or just some of the most recent samples to estimate the next p.d.f. [6].

The low number of bins (three) employed is a rough approximation of the true revenue p.d.f.. This level of detail is user-defined, along with the histogram classification rule set. The reduced length of the test set (i.e. one month) may not be enough to assume this setting as the best possible for this case study. Moreover, in more complex urban areas, it may be relevant to explore more complex p.d.f. approximations by determining which are the best parameter settings (i.e. ma, mi, μ and rule set) for each scenario. However, this discussion is not addressed in this paper.

In Fig. 6, it is possible to observe that the ensemble method has an accuracy $\geq 90\%$ in most stands. The busiest stands present a lower accuracy than expected. This behavior may indicate that there is a persistent error on this type of stand. However, a stand-based analysis on the algorithm's behavior is required to reach these conclusions.

Despite the limitations mentioned above, this work is only a fair proof of concept for using the demand numerical predictions in [5] to uncover the stands' profitability. Note that nearly 70% of the classification error results in a profitability class that is **lower** than the period's true label. This shows how reliable this methodology can be by being **cautious** to predict high-revenues. This approach may also benefit from employing a numerical ensemble instead (i.e.: as proposed in [5]). This work uses taxi stands as demand aggregation points. However, they may be replaced with spatial areas instead by dividing the urban area into non-overlapping regions [13]. The incremental properties of this time series may also be used to increase the prediction frequency [6].

VI. Final Remarks

This application paper proposes a novel technique to predict the short-term profitability of the taxi stand network spread throughout an urban area. The technique consists of typifying the services that will occur at each specific stand on a short term. The authors do so by predicting an **approximate revenue p.d.f.** at each stand by employing **time series analysis** techniques based on non-homogeneous Poisson processes [5]. Experiments conducted in a real-world case study demonstrated the validity of this concept by presenting a profitability classification accuracy of $\sim 74\%$. It provides a relevant contribution to the taxi-stand choice problem by predicting where the most profitable services will be requested, instead of looking solely at pick-up quantities.

The framework described along this paper is just a proof of concept of what can be done in the stand profitability prediction topic. There are mainly two issues to be mitigated in future work: (1) How can we determine the optimal parameter setting (i.e. ma,mi,μ and set of rules) for a given urban area? (2) How can we reduce the accuracy error at the busiest taxi stands? These are open research questions.

References

[1] Schaller Consulting, *The New York City Taxicab Fact Book*. Schaller Consulting, 2006.

[2] 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.

[3] J. Yuan, Y. Zheng, X. Xie, and G. Sun, "Driving with knowledge from the physical world," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011, pp. 316–324.

TABLE III: Profitability Prediction Evaluation.

Method	Accuracy	lower-prediction error	higher-prediction error	RMSE-bin1	RMSE-bin2	RMSE-bin3
Poisson Mean	73.63%	16.83%	9.54%	1.8243	1.3956	0.6052
Exponential Smoothing	71.57%	16.66%	11.77%	1.8944	1.4438	0.6406
ARIMA	70.91%	17.90%	11.19%	1.9441	1.4781	0.6285
Majority Class	65.19%	22.18%	12.64%	N/A	N/A	N/A
Ensemble	**73.99%**	**17.88%**	**8.13%**	N/A	N/A	N/A

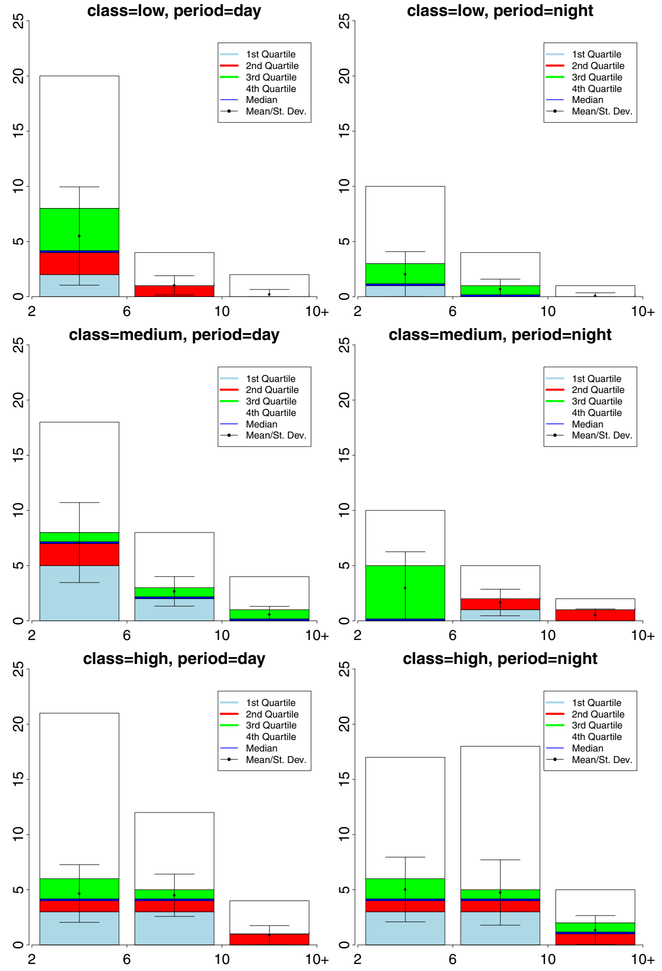

Fig. 5: Descriptive Statistics on each bin values for different periods and profitability classes at taxi stand 57.

[4] 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.

[5] L. Moreira-Matias, J. Gama, M. Ferreira, J. Mendes-Moreira, and L. Damas, "Predicting taxi-passenger demand using streaming data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 3, pp. 1393–1402, 2013.

[6] L. Moreira-Matias, J. Gama, M. Ferreira, J. Mendes-Moreira, and L. Damas, "On predicting the taxi-passenger demand: A real-time approach," in *Progress in Artificial Intelligence*, ser. LNCS. Springer, 2013, vol. 8154, pp. 54–65.

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

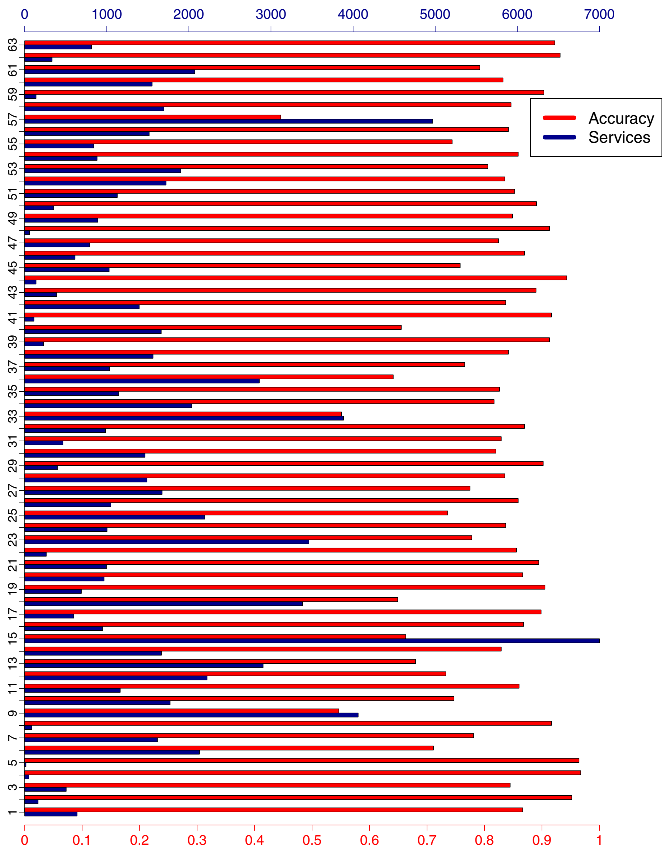

Fig. 6: Ensemble evaluation divided by stand. The grouped bars represent the accuracy (light red) and the number of services requested at each stand (dark blue).

[8] H. Chang, D. Park, S. Lee, H. Lee, and S. Baek, "Dynamic multi-interval bus travel time prediction using bus transit data," *Transportmetrica*, vol. 6, no. 1, pp. 19–38, 2010.

[9] H. Hu, Z. Wu, B. Mao, Y. Zhuang, J. Cao, and J. Pan, "Pick-up tree based route recommendation from taxi trajectories," in *Web-Age Information Management*. Springer, 2012, pp. 471–483.

[10] J. Powell, Y. Huang, F. Bastani, and M. Ji, "Towards reducing taxicab cruising time using spatio-temporal profitability maps," in *Advances in Spatial and Temporal Databases*. Springer, 2011, pp. 242–260.

[11] J. Gama and C. Pinto, "Discretization from data streams: applications to histograms and data mining," in *Proceedings of the 2006 ACM Symposium on Applied Computing*. ACM, 2006, pp. 662–667.

[12] K. Yeasmin and J. Rob, *Automatic Time Series Forecasting: The forecast Package for R*, 1999. [Online]. Available: http://oai.repec.openlib.org

[13] P. Castro, C. Zhang, D.and Chen, S. Li, and G. Pan, "From taxi gps traces to social and community dynamics: A survey," *ACM Comput. Surv.*, vol. 46, pp. 17:1–17:34, 2013.