A Two-Stage Model for Real-time Taxi Demand Prediction Using Data from the Web

Markou, Ioulia; Pereira, Francisco Camara; Rodrigues, Filipe

Published in:
Proceedings of the Transportation Research Board 98th Annual Meeting

Publication date:
2019

Document Version
Peer reviewed version

Transportation Research Record

A TWO-STAGE MODEL FOR REAL-TIME TAXI DEMAND PREDICTION USING DATA FROM THE WEB

--Manuscript Draft--

Full Title:	A TWO-STAGE MODEL FOR REAL-TIME TAXI DEMAND PREDICTION USING DATA FROM THE WEB
Manuscript Number:	19-04063R1
Article Type:	Presentation Only
Order of Authors:	Ioulia Markou
	Francisco Camara Pereira, Full Professor
	Filipe Rodrigues, Postdoc

A TWO-STAGE MODEL FOR REAL-TIME TAXI DEMAND PREDICTION USING DATA FROM THE WEB

Ioulia Markou
PhD Student, DTU Management Engineering, Transport DTU
Bygningstorvet 116B, DK-2800 Kgs. Lyngby
Tel: +45 45251515; Email: markou@dtu.dk

Francisco C. Pereira
Full Professor, DTU Management Engineering, Transport DTU
Bygningstorvet 116B, DK-2800 Kgs. Lyngby
Tel: +45 45251496; Email: camara@dtu.dk

Filipe Rodrigues
Assistant Professor, DTU Management Engineering, Transport DTU
Bygningstorvet 116B, DK-2800 Kgs. Lyngby
Tel: +45 45256530; Email: rodr@dtu.dk

Word count: 6,094 words text + 4 table x 250 words (each) = 7094 words

August 1st, 2018

ABSTRACT

Timely and accurate anticipation of phenomena in a variety of applications, such as economics, energy and transportation, is considered necessary for their smooth operation. Up to the present time, several efforts have been made in the transport sector to exploit information related to supply and demand for the formulation of accurate forecasting models. However, it is widely known that the model that could provide us with very accurate demand predictions under any circumstances has not yet been formed. Many factors in daily life, such as special events and traffic distruptions, overturn traffic system's balance, and the reliability of forecast models decreases significantly. The main focus of this research is the analysis, evaluation, and forecasting of prediction model's residuals in a real-time taxi demand forecasting framework. We comprise a deep learning architecture that is based on Fully-Connected dense layers. Publicly available taxi data from New York are explored, as well as semantic information combinations, that are typically neglected from modern techniques. The analysis focuses on two main areas, where significant fluctuations in demand are observed, due to popular venues located in the area. The performance of our proposed two-stage process with the inclusion of residuals' forecasts, is improved considerably.

Keywords: Time-series Forecasting, Deep Learning, Taxi demand, Special events, Semantic information, Topic modeling

INTRODUCTION

In general, mobility trends captured in complex transport systems consist of two basic components: utilitarian travel that mostly includes habitual behavior (e.g. commuting to work, weekly shopping) but also to a minor extent non-habitual needs (e.g. go to hospital, occasional shopping); and recreational travel, which comprises the human need for entertainment, social interaction and public expression. Efficient and effective intelligent transport systems should be able to take into consideration both of these factors for accurate demand predictions and better traffic management.

Current prediction approaches generally focus on capturing recurrent conditions, namely their seasonal spatial-temporal aspects (the "average" winter peak-hour Monday, in area X, with weather Y). The developed approaches can be successful for long-term planning applications or for modeling demand in non-eventful areas such as residential neighborhoods. However, in lively and dynamic areas where multiple special events take place, such as music concerts, sports games, festivals, parades and protests, these approaches fail to accurately model mobility demand precisely at times when it is needed - when the transport system of the area is under stress. The inability of the system to meet the new demand conditions emphasizes the need of good anticipatory capabilities which are capable to accept timely information on such phenomena.

Non-recurrent special events, such as concerts, sport games and demonstrations, are planned and largely advertised on the Web. An interesting fact is that it is much more likely to have citizens sharing their expectations/experiences about non-recurrent events than to talk about their daily commute. This plethora of information makes the Web an important tool for demand prediction and thus system's balance maintenance.

Previous studies have shown a strong correlation between number of public transport arrivals with the structured data mined from the Web *(1, 2)*. Namely, semi-structured information about events from announcements websites can be used as features for public transport arrivals. However, information contained on these websites is usually incomplete, noisy or missing, which makes it difficult to generalize. Going beyond this approach raises two challenges: which details about a scheduled event (time, type of event) are useful and how relevant information can be turned into model.

The aim of this study is the exploitation of information available on the internet for real-time demand prediction using a two-stage model. A particular emphasis will be given to venue areas, where several special events that are publicly disclosed on social media are hosted and attract many people. The proposed framework will be able to predict intervals of high demand that the average supply of the studied transport system (taxi services, uber etc.) cannot easily cover.

The paper is organized as follows. In the literature review section, we briefly review previous studies and discuss challenges related to information about events, traffic prediction and demand modeling using deep learning. In the methodology section we describe the proposed model and in the section of experiments we present the data used in its validation, as well as the tools for its implementation. The paper ends with the results and our conclusions.

LITERATURE REVIEW

Internet as a data source for special events

Internet, and more specifically the several social networking services that exist, has become a popular distribution outlet for users looking to share their experiences and interests on the Web. Taking as an example the Facebook, which has over 2.19 billion monthly active Facebook users

(Facebook MAUs) worldwide, it is clearly understood that the information derived from the above platforms, can undeniably help discerning explanations about observed real-world phenomena, such as non-habitual overcrowding scenarios.

Due to the importance of special events' impact in urban mobility, it is not surprising that they are a predominant part of transportation research. Fortunately, the Internet is rich in information about public special events. In an earlier work, Pereira et al. *(1)* compared an origin/destination (OD) prediction model based on public transport data with and without simple information obtained from the Internet, such as event type or whether the performer/event had a Wikipedia page. It was verified that such information could reduce the root mean squared error (RMSE) by more than 50% in each OD. In another study, Pereira et al. *(2)* presented a machine learning model that classifies aggregated crowd observations into explanatory components. After the identification of overcrowding hotspots in the city-state of Singapore, potential explanations from several event announcements websites were retrieved.

The internet is also a valuable source for other aspects of mobility research. For example, Twitter has been used for crisis management *(3, 4)*, urban management and planning (5), the analysis of different aspects of mobility (6) and the mobility characteristics of different nations *(7)*. Due to the complexity of the exploration of the open Web (e.g. using Google search), the use of internet data in transportation, however, is currently limited to manually defined sources and highly fine-tuned processes.

Demand Prediction for special events

Special events have a huge impact in urban mobility, regardless of their scale and type. Understanding their influence on the balance of a transport system is crucial for the development of reliable traffic management operations. For large-scale events (e.g. World cup, Formula One and Olympic games), best practices are already available for authorities to follow in order to manage these events and prepare for them well in advance *(8, 9)*. However, these manual approaches do not scale to the vast amount of smaller and medium-sized events that take place on large metropolitan areas on a daily basis. Despite their reduced scale, these events still have a significant impact in the transportation system *(2)*, especially when multiple co-occur. In these scenarios, common practice relies on reactive approaches rather than on planning *(10, 11)*. The demand prediction solution that we propose in this paper, takes into consideration event information that is automatically mined from the Web, and present itself with the potential for anticipating the effects of events and showing reliable tools for hotspot predictions in eventful areas.

Taxi demand has been the subject of several applications, since the related datasets are sufficiently detailed. The yellow and green taxi public dataset of New York City in particular, has been the subject of a lot of research. Morgul and Ozbay *(12)* present an empirical assessment of taxicab drivers' labor supply. Yang and Gonzales *(13)* identify locations and times of day where there is a mismatch between the availability of taxicabs and taxi service demand. Zhao et al. (*14)* use entropy and the temporal correlation of human mobility to measure the demand uncertainty at the building block level. They implemented three prediction algorithms to validate their maximum predictability theory. The importance of identifying hotspots, where demand is expected to be higher than the expected average demand is highlighted in the research of Markou et al. *(15)*. Through kernel density analysis, demand fluctuations were detected and analysed and significant deviations from the average day were correlated with disruptive event scenarios such as extreme weather conditions, public holidays, religious festivities, and parades. Finally, some other research

studies used this taxicab data to explore taxicab driver's airport pick-up decisions *(16)* or travel time variability analysis *(17).*

Deep models in transportation

Deep learning is evolving rapidly in solving problems that have resisted the best attempts of the artificial intelligence community for many years. It has proven to be able to find intricate structures in high-dimensional data, and thus it is an important tool in various applications in the domain of science *(18).* In the field of transportation and urban mobility there are already studies showing deep learning's successfullness.

Lv et al. *(19)* proposed a deep-learning-based traffic flow prediction method that takes into consideration the traffic flow features as learned by a stacked autoencoder model (SAE). Results comparison with more traditional approaches based on Support Vector Regression (SVR) and radial basis functions (RBFs) showed proposed method's superiority. Ma et al. *(20)* proposed a long short-term memory neural (LSTM) network for travel speed prediction. Their empirical results on data from Beijing indicate that LSTMs outperform other methods such ARIMA and SVR, which the authors justify with the ability of LSTMs to capture long-term dependencies over the time-series. A model with Mixture Density Networks (MDN) on top of LSTM was proposed by Xu et al. *(21).* In their approach, the city is previously divided in smaller areas and then the LSTM-based model is used to jointly predict the taxi demand for the next time-step in all the areas. Finally, the prediction of crowds' traffic in city's regions using a deep-learning based approach, called ST-ResNet, is presented by Zhang et al. (*22).* Experiments on two types of crowd flows in Beijing and New York City (NYC) demonstrate that the proposed method outperforms standard approaches such as ARIMA and vector auto-regressive models.

While the approaches described above demonstrate the potential of deep learning for transportation problems, none of these approaches consider the effect of events in order to improve their predictions. The deep learning approaches proposed in this paper aim at bridging this gap by focusing on event areas and showing that data fusion techniques can also help develop a model that is able to recognize conditions of greater prediction uncertainty, and thus improve the final demand forecasts.

METHODOLOGY

From previous research, we have already highlighted the importance of textual data for more accurate daily forecasts in event areas *(23).* The proposed neural network architectures lead to significant reductions in forecasting error using event information extracted from the Web. In this study, having at our disposal all possible information for future events, we present how data fusion can also be very useful on forecasting the error of our neural network architecture and thus on the even greater performance of our final model. Our focus is taxi demand prediction in real-time.

Model formulation

The proposed structure of the prediction model architecture includes two phases, (a) the training phase and (b) the test phase. In the first phase, only the first forecasting model (referred to as "Demand Prediction Model" – "DP-Model") is used, whose architecture is presented in the next subsection. The main objective of our DP-Model is to predict taxi demand based on the available historical data for the areas that we are interested in.

At the second phase, we use the forecasts we received at the end of the training phase, for the calculation of the DP-Model's forecast deviation from the actual demand. The obtained residuals that correspond to the previous timeframe are used as the training dataset of our second

deep learning model (referred to as "Error of Demand Prediction" - "EoDP-Model"), whose independent variables include the day of the week, day of month, topics and dummy variables that represent information about the presence of events before or after the predicted hourly demand. The objective of the EoDP-Model is the estimation of the demand prediction residuals based on the calculated residuals that the DP-Model attributed to that particular day of the month/week in the past, where an event was or was not scheduled.

At each stage we use separate training, validation and test sets and there are no time periods overlap between the two phases.

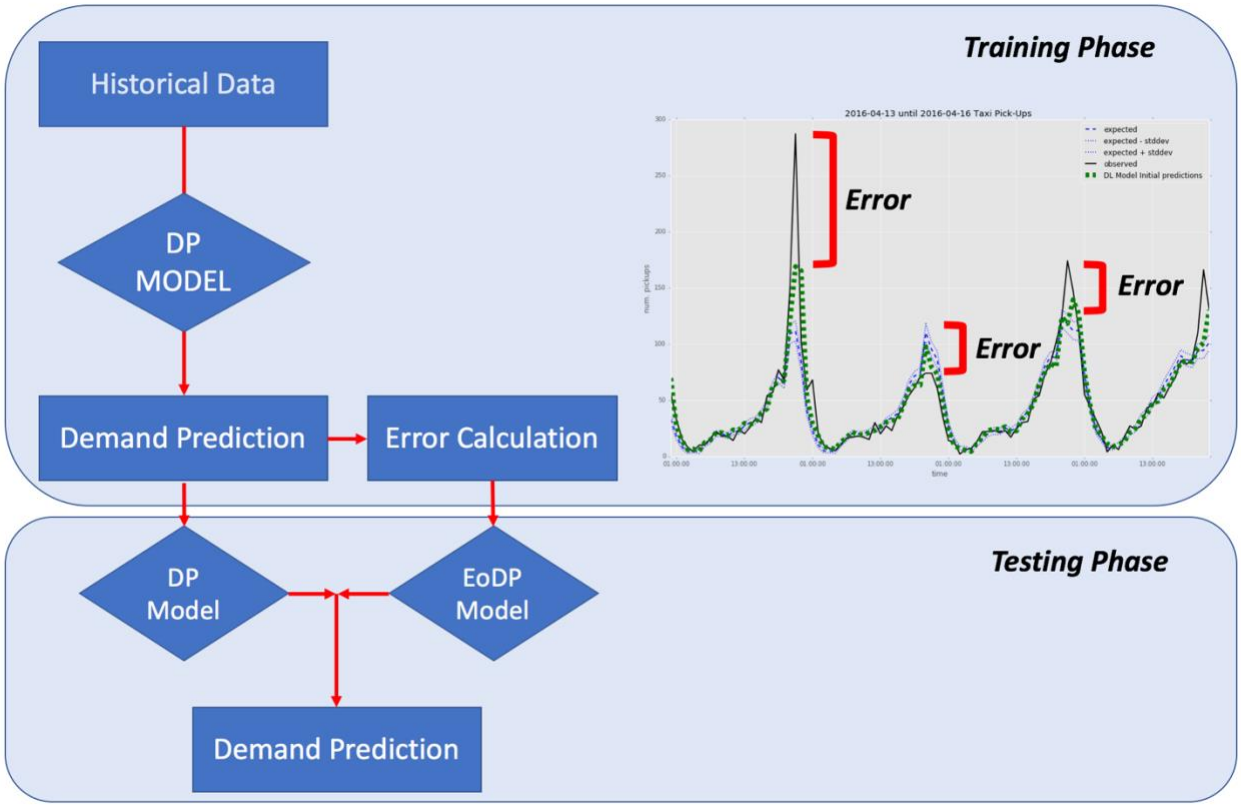

FIGURE 1 Proposed methodology

To elaborate:

- The training phase is subdivided into two time periods. During the first time period (A Period) the DP-Model runs independently and gives demand predictions for the second time period (B Period). These predictions are evaluated based on true demand values that were observed during the B period, and the vector of residuals (predicted – true_values) is obtained. That vector will be the training dataset for our EoDP-Model.
- the testing phase refers to a new time period (C Period), for which the DP-Model gives demand predictions and the EoDP-Model gives future residuals estimations. The final outcome is the sum of those predictions.

Time-series detrending

One of the most important steps in preparing time-series data for analysis is detrending *(25)*. For the particular case of urban mobility applications, such as traffic flow forecasting and taxi demand

prediction, there are obvious cyclic trends that result from daily commuting and other recurrent behaviors.

Before the configuration of initial modeling structure, we decided to remove any deterministic trends and focus our analysis on the remaining fluctuations. A simple, yet very effective way, of identifying these daily or weekly recurring patterns is by constructing a historical averages model, which computes the individual averages for each (hour of day, day of the week) pair based on historical data (from the train set only). The historical averages then represent a fixed recurring trend, which can be easily removed from the data.

Deep Learning Architecture

The data fusion architecture that makes use of fully-connected (FC) layers for modeling the time-series data is depicted in Figure 2. All the time-series information is provided as a flat input vector to the network in the form of lagged information. The network is fed with the values for the observations at $\{t, t-1, \cdots, t-L\}$ in a vector of size $L+1$, where $L+1$ corresponds to the number of lags. This vector is fed into a FC layer with 100-200 hidden units and hyperbolic tangent (*tanh*) activations, which also can receive additional inputs with other relevant information, such as the event details described in previous subsections. The output of this FC layer is then passed to a second FC layer with 50 units and *tanh* activations. We apply BatchNormalization *(24)* before every FC layer, Dropout between FC layers and we use regularization whenever necessary.

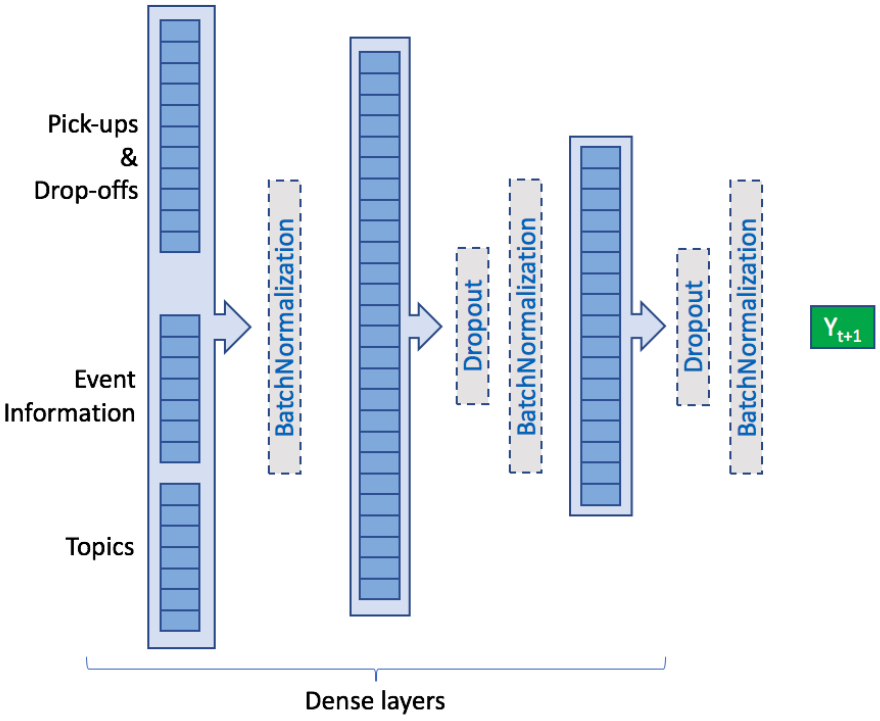

FIGURE 2 Proposed neural network architecture with FC layers (1st DL Model).

The idea is that the output of the last FC layer corresponds to a latent vector representation that encodes all the necessary information form the time-series and other relevant inputs. From this latent vector representation, we will finally produce a prediction for $t+1$ using a dense layer. The

final prediction is obtained by adding back the removed recurrent trend (based on the historical average) to the output of the neural network.

Text data pre-processing

Generally, textual data mined from the Web is difficult to process in its original state. Specific pre-processing steps are usually required in order to make it more amenable to learning methods, and more specifically to the topic modelling stage that will follow. Therefore, we follow a simple conventional text-processing pipeline consisting of:

- HTML tag removal.
- Lowercase transformation for words' variability restriction purposes.
- Tokenization, a tool that divides a sequence of characters into pieces of tokens.
- Lemmatization for inflectional endings removal, and words return to their base form (lemma).
- Stopwords and very frequent words removal, which typically do not bring any additional useful information.
- Removal of words that appear only once in the whole dataset.

Topic Modeling

A considerable amount of important information about the planned events that we have at our disposal is in textual form. Their description, title or comments on the website hosting the announcement include many useful details about the type of event and, by extension, its popularity. Therefore, in order to build a forecasting model that takes into consideration text description and historical taxi demand data, we need to convert such data into a proper representation that our deep learning algorithm can understand. However, the dimensionality of the deep learning model will be increased beyond reasonable if we explicitly include the text, word by word. Natural language is rich in synonymy and polysemy, different announcers and locations may use different words, besides it is not always obvious which words are more "relevant". Topic modeling is the research topic that focuses on covering these weaknesses.

The approach of topic modeling is to represent a text document as a finite set of *topics*. These topics correspond to sets of words that tend to co-occur together rather than a single word associated with a specific topic. For example, a rock festival textual description could have a weight w_1 assigned to topic 1 (e.g. words related to concerts in general), w_2 of topic 2 (e.g. words related to festivals), w_3 of topic 3 (e.g. words related to the venue descriptions) and so on. In particular, we use a specific technique that is called Latent Dirichlet Allocation (LDA). For the readers that are familiar with Principal Components Analysis (PCA), there is a simple analogy: PCA re-represents a signal as a linear combination of its eigenvectors, while LDA re-represents a text as a linear combination of topics. In this way, we reduce the dimensionality from the total number of different words of a text to the number of topics, typically very low. Each document is represented as a distribution over topics, and each topic is a distribution over words.

LDA's generative process includes the following steps:

- Draw a topic β_k from $\beta_k \sim Dirichlet(\eta)$ for k = 1...K
- For each document *d:*
  - Draw topics proportions ϑ_d such that $\vartheta_{d\,k} \sim Dirichlet(a)$
  - For each word $w_{d,n}$:
    - Draw topic assignment $z_{d,n} \sim$ Multinomial(ϑ_d)

- Draw word $w_{d,n} \sim$ Multinomial($\beta_{Zd,n}$)

The parameters a and η are hyperparameters that indicate respectively the priors on per-document topic distribution and per-topic word distribution, respectively. Thus, $w_{d,n}$ are the only observable variables, all the others are latent in this model. For a set of D documents, given the parameters a and η, the joint distribution of a topic mixture ϑ, word-topic mixtures β, topics z, and a set of N words is given by:

$$p(\vartheta,\beta,z,w|\alpha,\eta) = \prod_{k=1}^{K} p(\beta_k|n) \prod_{d=1}^{D} p(\vartheta_d|a) = \prod_{n=1}^{N} (p(z_{d,n}|\vartheta_d)p(w_{d,n}|\beta_k, k = z_{d,n})) \quad (1)$$

Broadly speaking, the training task is to find the posterior distribution of the latent variables (the per-document topic proportions ϑ_d, the per-word topic assignments $z_{d,n}$ and the topics β_k) that maximize this probability.

For further details concerning LDA's generative process please refer to the original article of David Blei and colleagues *(26)*.

The parameter that we mainly focused in this study is the number of topics. We tested a range of values between 5 and 30, and we empirically concluded that the value of 10 yielded the best model results. With 10 topics we are able to capture all kinds of events included in our event database, and we also narrow down possible equivocal topics that could deteriorate our results. The other parameters, the a and η priors, were kept as default. The LDA results for Barclays Center are presented in Table 1.

TABLE 1 LDA Results

Topic	No. Events with θd > 0.8	Popular Words
Topic_1	24	ice, disney, present, magic, new
Topic_2	72	basketball, championship, atlantic, game, tournament
Topic_3	10	show, artist, box, office, special
Topic_4	32	music, atlantic, championship, basketball, game
Topic_5	22	game, marriot, corporate, bridge, hotel
Topic_6	10	train, service, islander, view, time
Topic_7	34	tour, album, show, meet, up
Topic_8	12	circus, family, out, space, earth
Topic_9	12	dinner, reservation, jay, menu, restaurant
Topic_10	42	champion, game, group, boxing, hoop

EXPERIMENTS

In this section, we demonstrate the hypothesis that information about events is significant in real-time taxi demand prediction in the vicinity of special event venues. The inclusion of information about the occurrence of planned event, allows a better understanding of demand fluctuations, as well as the restriction of final forecasts' margin of error. Our approach is evaluated in two event areas in New York City (NYC) and the proposed data fusion methodology was implemented in Keras *(27)*.

Dataset and case studies

Our base dataset consists of 1.1 billion taxi trips from New York, distributed by technology providers of authorized under the Taxicab & Livery Passenger Enhancement Programs (TPEP/LPEP) and were made publicly available by the NYC Taxi and Limousine Commission (TLC). We use taxi data from 1/1/2013 through 6/30/2016, which includes around 600 million taxi trips after data filtering. The dataset specifies for each drop-off and pick-up event the GPS location and the time-stamp.

Based on this data, we looked at a list of the top venues in NYC and selected the two venues for which more complete event records were available online: the Barclays Center and Terminal 5. The first venue is located in the heart of Brooklyn and it is the state-of-the-art home of the NBA's Brooklyn Nets and the NHL's New York Islanders. It is one of the most popular facilities in the New York metropolitan area because it hosts many sold-out concerts, conventions and other sporting and entertainment events. It is ranked top five globally in 2015 for gross revenue and attendance by Billboard and Venues Today *(28)*. On the other hand, the Terminal 5 is a 3-floor venue that regularly hosts concerts with many different audiences and that is located in the heart of Manhattan. Given the geographical coordinates of these two venues, we selected all the taxi pickups that took place within a bounding box of ±0.003 decimal degrees (roughly 500 meters) to be our study areas (Figure 3).

The individual records that fall within the boundaries described above were grouped in a time-series of hourly counts. Our goal is to predict the taxi demand of the area at the next hour, considering the demand from previous records, as well as event information extracted from the Web. In this way, stakeholders, such as companies like Uber and taxi operators, can have a clear image of demand in the near future, thus allowing them to better organize their fleet. Precise next-hour demand forecasts allow those companies' fleet to become more efficient as routes become targeted and balanced with the demand.

Regarding the event data, it was extracted automatically from the Web using either screen scrapping techniques or Application Programming Interfaces (API's). For the Barclays Center, the event information was scrapped from its official website, since it maintains a very accurate and detailed calendar. We collected a total of 751 events since its inauguration in late 2012 until June 2016. As for the Terminal 5, we used the Facebook API to extract 315 events from its official page, for a similar time period. In both cases, the event data includes event's title, date, time and description.

FIGURE 3 Map of the two studied areas. Map data © OpenStreetMap contributors

Experimental Setup

The taxi dataset includes records of trips from 2013 and we created separate training and test sets for each stage of our methodology (Figure 4). More specifically, we selected the first three years (January 2013 – December 2015) as our training set and the first 5 months of 2016 as our test set for the first phase of our methodology with the "DP-Model". For the validation process, we separated 20% of our training set using the automatic tools of Keras.

At the second stage, where the "EoDP-Model" is introduced, we extend the training set of the "DP-Model" to May 2016, and we use the last month of our dataset (June 2016) for testing. The first five months of 2016 are used as the training dataset of "EoDP-Model", since the calculated error values from the first stage correspond to this timeframe.

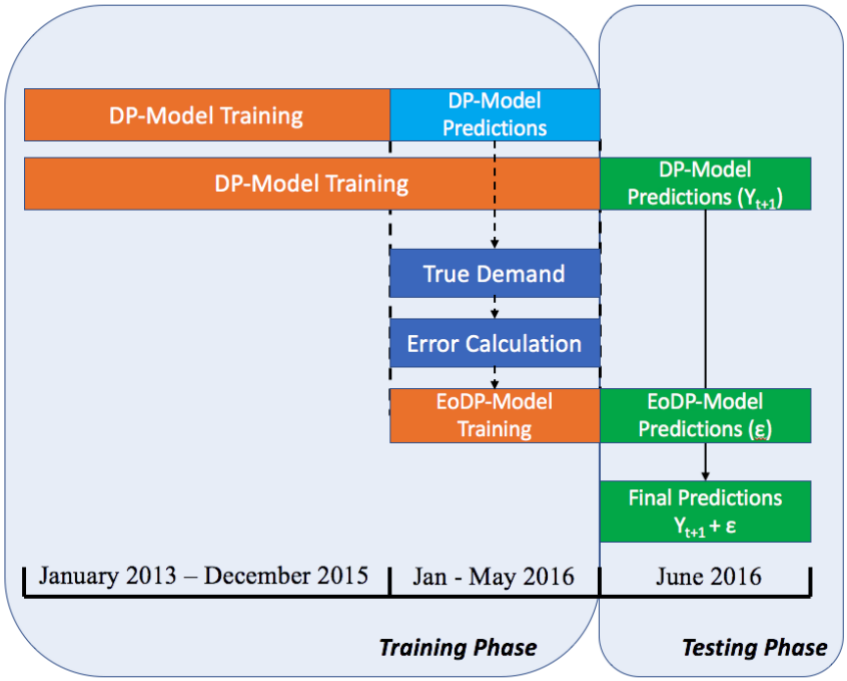

FIGURE 4 Experimental Setup Depiction.

Training Phase

In order to evaluate the contribution of the different sources of information, we perform an incremental analysis of the proposed deep learning architecture. We start with only the part of the network that is responsible for modeling the time-series data and we keep adding components to the network until the full model depicted in Figure 2 is obtained. Therefore, we start with a model that only takes the lagged pickup observations (referred to as "P") as input and move to models that also include: drop-off lags (denoted "P+D"), information about the presence of events ("P+D+E") and finally, the full model that also considers events' topics ("P+D+E+T").

For models' performance validation and comparison, we will use the mean absolute error (MAE), the root-mean-square error (RMSE) and the coefficient of determination (R2).

After the forecasts are obtained, the deviation of each measurement from the true value is calculated. The new vector will be used as the dependent variable of the EoDP-Model at the testing phase of the proposed methodology.

Testing Phase

At this stage, we implement a parallel training of our models. For the DP-Model, the training dataset is expanded by 5 months (Jan-May 2016) and for the EoDP-Model we use the vector of residuals that correspond to the same time period Jan-May 2016.

The evaluation of our approach starts with the day of month and day of week as EoDP-Model's independent variables (denoted "d+m") and we add event information ("d+m+E") and events' topics ("d+m+E+T") in the subsequent steps. The EoDP-Model's architecture makes use of fully-connected (FC) layers as as in the DP-Model model. The only difference is the composition of the first dense layer, since we do not have as inputs the pickups and drop-offs lags, but the day of week and day of month.

For the step-by-step and detailed analysis of the contribution of each parameter, we begin with the study of the basic models and gradually add more parameters. Therefore, at first, we check the accuracy of the models using only pickup lags (for the DR-Model) and the day of week and day of month variables (for the EoDP-Model), and afterwards we add more variables that could improve method's performance.

Besides these baselines, the proposed approach is further compared with another popular method from the state of the art for time-series forecasting, the Linear Regression (LR). Its simplicity and interpretability were the criterias for its selection.

RESULTS

Table 2 shows the results of demand predictions using *only* the DP-Model for June 2016. They will form our baseline for the evaluation and comparison of that simple architecture of a single model with the proposed two-step approach that we will implement afterwards.

From the initial results we understand that in the case of linear regression, information about scheduled events plays an influential role. For Barclays center, the type of event contributes more to the results' improvement, while for Terminal 5 the start and end time of an event seems to be more determinant. This conclusion can be also justified by the fact that the popular venue of Brooklyn hosts concerts, conventions and other sporting and entertainment events, which attract a different number of people each time, and demand fluctuations can also be different because of that. Topic modeling captures the event categories that the venue hosts, therefore its contribution is clear on the final accuracy of the model. On the other hand, Terminal 5 hosts mostly music events and and audience attendance can be estimated satisfactorily with the start and end time of each event.

In the case of deep learning, it is obvious that event information does not appear to have any significant effect on the results, when it is included directly to the demand forecasting model. Pickup and drop-off lags seem enough for model's best possible performance. It remains to be seen, if by using the proposed architecture with the EoDP-Model, the results will be changed.

Table 3 shows the performance of our two-step architecture for Barclays Center in June 2016, namely the same time period that Table 1 results refer to. In these measurements, both models are used, based on the methodology described in the methodology section.

We can see from the final scores that the introduction of a demand error forecasting model contributes significantly to the reduction of the final forecasting error. It is notworthy to mention that the EoDP-Model with event information has the greatest impact. In both cases (DL and LR models) the error forecasting model contributes positively.

TABLE 2 Demand prediction using only the DP-Model

	BARCLAYS CENTER			TERMINAL 5		
	MAE	**RMSE**	**R2**	**MAE**	**RMSE**	**R2**
LR-P	7,989	13,266	0,762	7,966	12,203	0,770
LR-P+D	7,437	11,817	0,811	7,159	10,755	0,821
LR-P+D+E	7,527	12,265	0,796	**7,134**	**10,487**	**0,830**
LR-P+D+E+T	**7,187**	**11,188**	**0,831**	7,163	10,558	0,828
DL-P	8,081	13,393	0,757	7,954	12,252	0,768
DL-P+D	7,671	12,121	0,801	7,079	10,556	**0,828**
DL-P+D+E	7,630	12,202	0,799	7,094	10,743	0,822
DL-P+D+E+T	7,625	12,106	**0,802**	7,151	10,736	0,825

For the deep learning models, we can observe that with the second model the final accuracy of the predictions is considerably increasing. The MAE is reduced by 5,2% and the RMSE by 8,04%. Using the DL method with pickups and drop-off lags and event information, it leads to an overall MAE reduction of 6% and 12% in RMSE. Therefore, it is obvious that model predictions can be further improved using the proposed architecture.

TABLE 3 Demand prediction using DP-Model and EoDP-Model (Barclays Center)

	MAE			RMSE			R2		
	d+m	**d+m+E**	**d+m+ E+T**	**d+m**	**d+m+E**	**d+m+ E+T**	**d+m**	**d+m+E**	**d+m+ E+T**
DL-P	8,080	**7,955**	**7,902**	13,288	**12,496**	**12,049**	0,761	**0,789**	**0,804**
DL-P+D	7,689	**7,365**	**7,598**	12,212	**11,152**	**11,711**	0,798	**0,832**	**0,814**
DL-P+D+E	7,672	**7,293**	**7,506**	12,157	**11,133**	**11,884**	0,800	**0,832**	**0,809**
DL-P+D+E+T	7,632	**7,355**	7,717	12,080	**11,302**	12,194	0,803	**0,827**	0,799
LR-P	8,034	**7,743**	8,067	13,293	**12,177**	14,250	0,761	**0,799**	0,725
LR-P+D	7,439	**7,072**	7,379	11,817	**10,714**	12,577	0,811	**0,845**	0.786
LR-P+D+E	7,528	**7,074**	7,417	12,260	**10,798**	12,719	0,797	**0,842**	0,781
LR-P+D+E+T	7,185	**7,105**	7,484	11,168	**10,775**	13,023	0,831	**0,843**	0,770

For Linear Regression, significant differences appear only using the "**d+m+E"** EoDP-Model. The MAE is decreased by 1,6%, which is also considered important, since the previous model (Table 2 results) was already fairly accurate.

Moving to the Terminal 5 study area, Table 4 shows the obtained results. In this case, the results are not as clear as in the previous study area. The positive contribution of the EoDP-Model using the DL method appears only in the "d+m+E" case, where the R^2 score of an already good forecasting model is still increased by 1,7%.

It seems that the mobility patterns around this venue, when a special event is organized, are less predictable than in Barclays Center. This is probably due to its location, which is in a very central area of Manhattan, where we can also locate some other popular venues, bars and restaurants that citizens prefer to visit daily. Consequently, the observed demand fluctuations in

this area are directly affected by other parameters which are not considered in this study, and therefore can not be predicted.

TABLE 4 Demand prediction using DP-Model and EoDP-Model (Terminal 5)

	MAE			RMSE			R2		
	d+m	d+m+E	d+m+ E+T	d+m	d+m+E	d+m+ E+T	d+m	d+m+E	d+m+ E+T
DL-P	8,162	**8,004**	8,229	12,444	**11,989**	13,681	0,761	**0,778**	0,711
DL-P+D	7,280	**7,133**	7,194	10,695	**10,358**	10,827	0,823	**0,834**	0,819
DL-P+D+E	7,067	**6,975**	7,624	10,613	**10,316**	10,035	0,826	**0,836**	0,834
DL-P+D+E+T	7,360	7,319	7,438	10,986	11,056	11,703	0,814	0,811	0.788
LR-P	8,072	8,072	8,112	12,257	12,257	12,385	0,768	0,768	0,763
LR-P+D	7,188	7,188	7,246	10,752	10,752	10,884	0,821	0,821	0,817
LR-P+D+E	7,158	**7,158**	7,241	10,474	**10,474**	10,91	0,830	**0,831**	0,816
LR-P+D+E+T	7,186	7,186	7,306	10,544	10,544	11,258	0,828	0,828	0,804

CONCLUSION

We demonstrated that using online information, we can improve the quality of taxi demand prediction even in scenarios where the transport system is under stress. We combined information extracted from the web with time-series data to formulate our two-step approach with tow predictive models that capture in real-time future demand in event areas. This is typically a challenging case for transport planning since special events originate high variance in demand. Taxi demand is correlated with many parameters of underlying information and currently, most taxi centers rely on formal processes and manual work for a fleet organization and taxi distribution. Even the more advanced new services, like Uber or Lyft, still face great challenges in terms of demand prediction (shown by price surges and variations thereof). Our results show a second model that predicts the forecasting error of the main model is able to further improve the final predictions. Information about events from the Web contributes decisively to the ultimate accuracy of the proposed methodology. Hence, besides the value of event information and time series data, our empirical results also highlight the need for accounting for the effect of events when modeling mobility demand.

In future work, we aim at exploring the impact of spatio-temporal interactions on taxi demand prediction. The development of a city-wide spatio-temporal model that accounts for information about all the events that take place across the city could be a generalization potential of this methodology.

AUTHOR CONTRIBUTION STATEMENT

The authors confirm contribution to the paper as follows: Study conception and design: I. Markou, F. C. Pereira, data collection: I. Markou, F. Rodrigues; analysis of results: I. Markou; draft manuscript preparation: I. Markou. All authors reviewed the results and approved the final version of the manuscript.

REFERENCES

1. Pereira C., Francisco, Filipe Rodrigues, and Moshe Ben-Akiva. Internet as a sensor: a case study with special events. *In 91st Transportation research board annual meeting* No. 12-3365. 2012.

2. Pereira C., Francisco, Filipe Rodrigues, Evgheni Polisciuc, and Moshe Ben-Akiva. Why so many people? explaining nonhabitual transport overcrowding with internet data. *IEEE Transactions on Intelligent Transportation Systems* 16, no. 3 (2015): 1370-1379.
3. Thom, Dennis, Harald Bosch, Steffen Koch, Michael Wörner, and Thomas Ertl. Spatiotemporal anomaly detection through visual analysis of geolocated twitter messages. *Visualization Symposium (PacificVis), 2012 IEEE Pacific*, pp. 41-48. IEEE, 2012.
4. Sakaki, Takeshi, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pp. 851-860. ACM, 2010.
5. Frias-Martinez, Vanessa, Victor Soto, Heath Hohwald, and Enrique Frias-Martinez. Characterizing urban landscapes using geolocated tweets. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pp. 239-248. IEEE, 2012.
6. Cheng, Zhiyuan, James Caverlee, Kyumin Lee, and Daniel Z. Sui. Exploring millions of footprints in location sharing services. *ICWSM* 2011 (2011): 81-88.
7. Hawelka, Bartosz, Izabela Sitko, Euro Beinat, Stanislav Sobolevsky, Pavlos Kazakopoulos, and Carlo Ratti. Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science* 41, no. 3 (2014): 260-271.
8. Dunn Jr, Walter M., Steven P. Latoski, and Elizabeth Bedsole. Planned Special Events: Checklists for Practitioners. No. FHWA-HOP-06-113. 2006.
9. Coutroubas, F., and Tzivelou N. Public transport planning for the greatest event: the 2004 Olympic Games."In *Proceedings of the European Transport Conference (ETC)* France. 2003.
10. Fuhs, Chuck, and Parsons Brinckerhoff. *Synthesis of Active Traffic Management Experiences in Europe and the United States*. No. FHWA-HOP-10-031. United States. Federal Highway Administration, 2010.
11. Kuppam, Arun, Rachel Copperman, Thomas Rossi, Vladimir Livshits, Lavanya Vallabhaneni, Ted Brown, and Kathy DeBoer. Innovative methods for collecting data and for modeling travel related to special events. *Transportation Research Record* 2246, no. 1 (2011): 24-31.
12. Morgul, Ender Faruk, and Kaan Ozbay. "Revisiting labor supply of new york city taxi drivers: Empirical evidence from large-scale taxi data." In *Transportation Research Board 94th Annual Meeting*, no. 15-3331. 2015.
13. Yang, Ci, and Eric J. Gonzales. "Modeling taxi demand and supply in New York City using large-scale taxi GPS data." In *Seeing Cities Through Big Data*, pp. 405-425. Springer, Cham, 2017.
14. Yang, Ci, and Eric J. Gonzales. Modeling taxi demand and supply in New York City using large-scale taxi GPS data. In *Seeing Cities Through Big Data*, pp. 405-425. Springer, Cham, 2017.
15. Markou, Ioulia, Filipe Rodrigues, and Francisco C. Pereira. Use of Taxi-Trip Data in Analysis of Demand Patterns for Detection and Explanation of Anomalies. *Transportation Research Record: Journal of the Transportation Research Board* 2643 (2017): 129-138.
16. Yazici, M. Anil, Camille Kamga, and Abhishek Singhal. A big data driven model for taxi drivers' airport pick-up decisions in new york city. In *Big Data IEEE International Conference.* pp. 37-44. IEEE, 2013.

17. Kamga, Camille. Temporal and weather-related variation patterns of urban travel time: Considerations and caveats for value of travel time, value of variability, and mode choice studies. *Transportation Research Part C: Emerging Technologies* 45 (2014): 4-16.
18. LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature* 521, no. 7553 (2015): 436.
19. Lv, Yisheng, Yanjie Duan, Wenwen Kang, Zhengxi Li, and Fei-Yue Wang. Traffic flow prediction with big data: A deep learning approach. *IEEE Trans. Intelligent Transportation Systems* 16, no. 2 (2015): 865-873.
20. Ma, Xiaolei, Zhimin Tao, Yinhai Wang, Haiyang Yu, and Yunpeng Wang. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data."*Transportation Research Part C: Emerging Technologies* 54 (2015): 187-197.
21. Xu, Jun, Rouhollah Rahmatizadeh, Ladislau Bölöni, and Damla Turgut. Real-Time Prediction of Taxi Demand Using Recurrent Neural Networks. *a) A* 50, no. 60 (2017): 70.
22. Zhang, Junbo, Yu Zheng, and Dekang Qi. Deep Spatio-Temporal Residual Networks for Citywide Crowd Flows Prediction. In *AAAI*, pp. 1655-1661. 2017.
23. Rodrigues Filipe, Markou Ioulia, and Francisco C. Pereira. Combining time-series and textual data for taxi demand prediction in event areas: a deep learning approach. *Information Fusion (accepted for publication).*
24. Ioffe, Sergey, and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* (2015).
25. Wu, Zhaohua, Norden E. Huang, Steven R. Long, and Chung-Kang Peng. "On the trend, detrending, and variability of nonlinear and nonstationary time series." *Proceedings of the National Academy of Sciences* 104, no. 38 (2007): 14889-14894.
26. Blei, David M., Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *Journal of machine Learning research* 3, no. Jan (2003): 993-1022.
27. Chollet, François, Keras (2015).
28. Barclays Center (2018, October 24). Retrieved from *https://www.barclayscenter.com/center-info/about-us*