Multi-step ahead prediction of taxi demand using time-series and textual data

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

Published in:
Transportation Research Procedia

Link to article, DOI:
10.1016/j.trpro.2019.09.094

Publication date:
2019

Document Version
Publisher's PDF, also known as Version of record

Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 41 (2019) 540–544

Transportation Research Procedia

www.elsevier.com/locate/procedia

International Scientific Conference on Mobility and Transport
Urban Mobility – Shaping the Future Together
mobil.TUM 2018, 13-14 June 2018, Munich, Germany

Multi-step ahead prediction of taxi demand using time-series and textual data

Ioulia Markou[a]*, Filipe Rodrigues[a], Francisco C. Pereira[a]

[a]*Technical University of Denmark (DTU), Bygning 116B, 2800 Kgs. Lyngby, Denmark*

Keywords: Time series forecasting; Textual data; Taxi demand; Special events

1. Introduction

Modelling urban mobility and understanding what drives the travel behavior of people is the key research topic for developing effective and efficient intelligent transportation systems that adapt to the travel demand. Typical forecasting approaches focus only on capturing recurrent mobility trends that relate to routine behaviors (Krygsman et al., 2004), and on exploiting short-term correlations with recent observation patterns (Moreira-Matias et al., 2013 and Van Oort et al., 2015). While this type of approaches can be successful for long-term planning applications or for modelling demand in non-eventful areas such as residential neighborhoods, in lively and highly dynamic areas that are prone to the occurrence of multiple special events, such as music concerts, sports games, festivals, parades and protests, these approaches fail to accurately model mobility demand (Pereira et al., 2015). As we move towards the deployment of autonomous vehicles, understanding and being able to anticipate mobility demand becomes crucial, especially in shared-mobility scenarios, as this allows for properly managing fleets and increasing user-satisfaction.

* Corresponding author. Tel.: +45 4525 1515.

Email addresses: markou@dtu.dk (Ioulia Markou), rodr@dtu.dk (Filipe Rodrigues), camara@dtu.dk (Francisco C. Pereira)

10.1016/j.trpro.2019.09.094

1.1. Taxi Demand Prediction

In situations where the traffic system is under stress (e.g. music concerts, sport games and political rallies), taxi-calling platforms, such as Uber, Grab and Beat are becoming increasingly popular, because they can efficiently facilitate resource allocation. Through their application, passengers are able to call or pre-order a taxi, even when they are located in an area where it is very hard to find a driver. Therefore, this trend proves that there is a tremendous need for better taxi fleet organization and taxi distribution from a taxi center, according to the demand of an entire city (Chan et al., 2016).

Several methods have been proposed to predict taxi demand, including probabilistic models (Yuan et al., 2011) neural networks (Xu et al., 2017) and time series modeling (Davis et al., 2016 and Moreira-Matias et al., 2013). The task is challenging because it is correlated with many parameters of underlying information. Currently, the general practice is to rely on formal processes and manual work. For very big events, such us the Olympic games or football world cup matches, the event organizers engage with operators and authorities to meet the enormous demand. For smaller events though, this task is labour-intensive and even with a list of events, their impact is hard to estimate. A timely and accurate notion of demand impact is accordingly needed in order to design adequate system changes and to disseminate appropriate information to the public.

2.1. Internet as a data source for special events

The progress and development of information technology has accelerated the growth of mobile technology mediated environments. Along with the evolution of the Internet, especially the growing of mobile data due to more and more mobile devices and applications, the information shared by each of us publicly is relatively increasing. Consequently, the Internet has become a valuable source for mining information about mobility in the city. Through popular websites and social platforms such as Facebook, Twitter, Wikipedia, eventful.com Foursquare, etc., it is possible nowadays to collect information about popular events that happened in the past, as well as information for events planned in the near future.

However, most of this information is typically in the form of unstructured natural-language text. Solving this cross-domain data fusion challenge then becomes key for understanding the mobility demand patterns that are caused by events, and also for addressing the general class of problems where text data from the Web can provide the context for explaining some of the patterns that are observed in time-series data. These not only are quite ubiquitous and cover various research fields, but they are becoming increasingly relevant as people share more and more information online. Popular examples include the use of text data from online social media to help predict financial time-series (e.g. stock markets (Tang et al., 2009 and Si et al., 2013)) and opinion polls (O'Connor et al., 2010).

This research aims at exploring machine learning architectures for combining time-series and textual data for multi-step ahead prediction of mobility demand in event areas. Specifically, we focus on the problem of taxi demand prediction, although the proposed methodology is applicable to other transportation modes as well. We empirically show the value of modelling the textual information associated with the events, and that the proposed machine learning approaches are able to improve their predictions significantly by combining information from different sources and formats.

2. Dataset and case studies

The base dataset for our experiments consists of 1.1 million taxi trips from 355 New York (January 2009 to June 2016) that were made publicly available by the NYC Taxi & Limousine Commission (TLC, 2017). Based on this data, we then looked at a list of the top venues in NYC (Best concert venues in NYC, 2017) and selected the two venues for which more complete event records where available online: the Barclays Center and Terminal 5. Located in Brooklyn, the Barclays Center is modern multi-purpose arena with 360 18.000 seats that regularly hosts major musical performances and serves as the new home of the NBA's Brooklyn Nets. On the other hand, the Terminal 5 is a 3-floor venue that regularly hosts concerts with many different audiences and is located in the heart Manhattan. Figure 1

shows a map of these areas. 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. The filtered taxi trip records are additionally aggregated by hour.

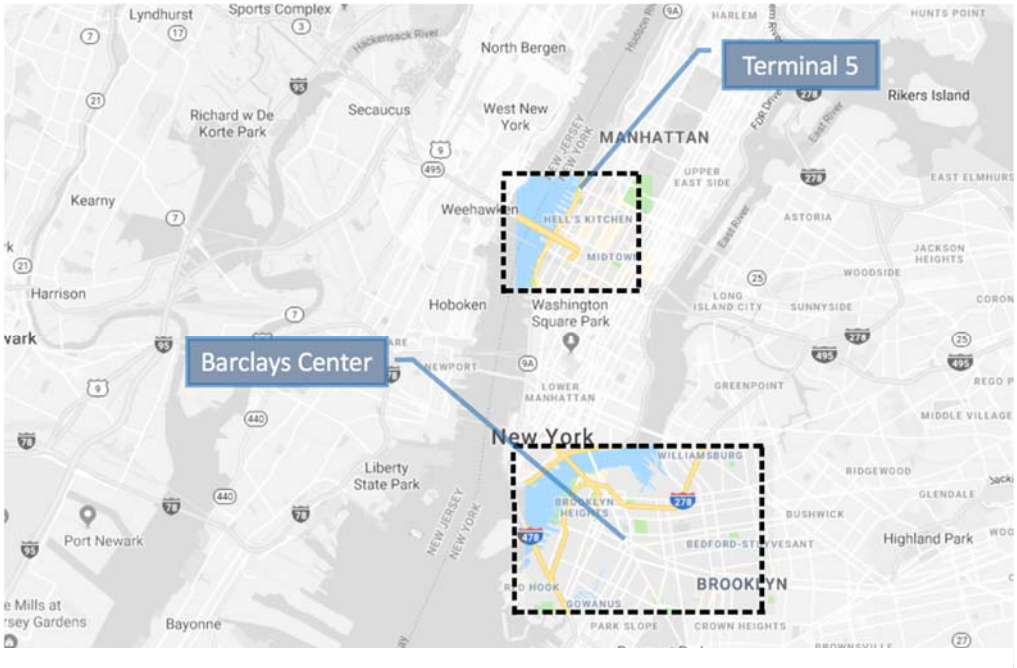

Fig. 1. Map of the two study areas.

Regarding the event data, it was extracted automatically from the Web using either screen scrapping or API's. For the Barclays Center, the event information 385 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 for a similar time period. In both cases, the event data includes event title, date, time and description.

3. Multi-step ahead prediction

Based on the data from the two case-study areas described in the previous section, we created separate training and test sets. The goal is to predict the number of taxi pickups in an area for the next 24 hours given the data from the previous days, weather data and event information extracted from the Web.

We developed two different multi-step prediction scenarios. The first one uses historical data as its training set and predicts taxi demand for the next 24 time slots (Fig. 2). Each time slot prediction is independent from the rest. The second scenario consists of two phases: in the first phase, the model predicts the taxi demand of the following time slot (y_{t+1}), and in the second phase it updates its training vector with the provided prediction and continues with the demand forecasting of the next time slot (y_{t+2}) (Fig. 3). The proposed approach is implemented using two popular methods from the state of the art for time-series forecasting: linear regression and Gaussian processes (GPs). All combinations of event and no-event days are studied, and the corresponding predictions are taken into consideration.

Through the implemented set-ups, the importance of the various parameters introduced in the model is examined. Great emphasis is given on the importance of textual data in the form of topics, for the accuracy of our predictions. The outputs of our implemented scenarios are compared with other corresponding forecasting models and their contribution to them is being studied.

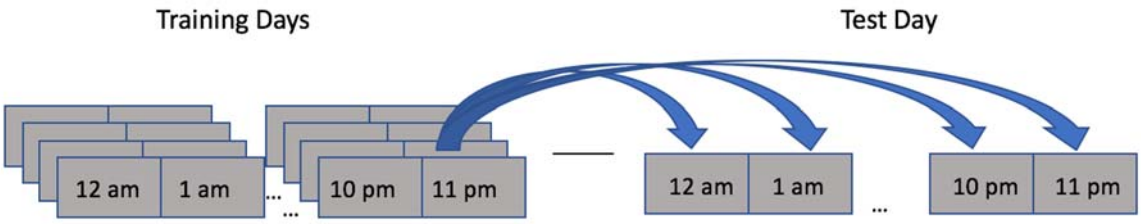

Fig. 2. Scenario 1 - Multi-step predictions without update

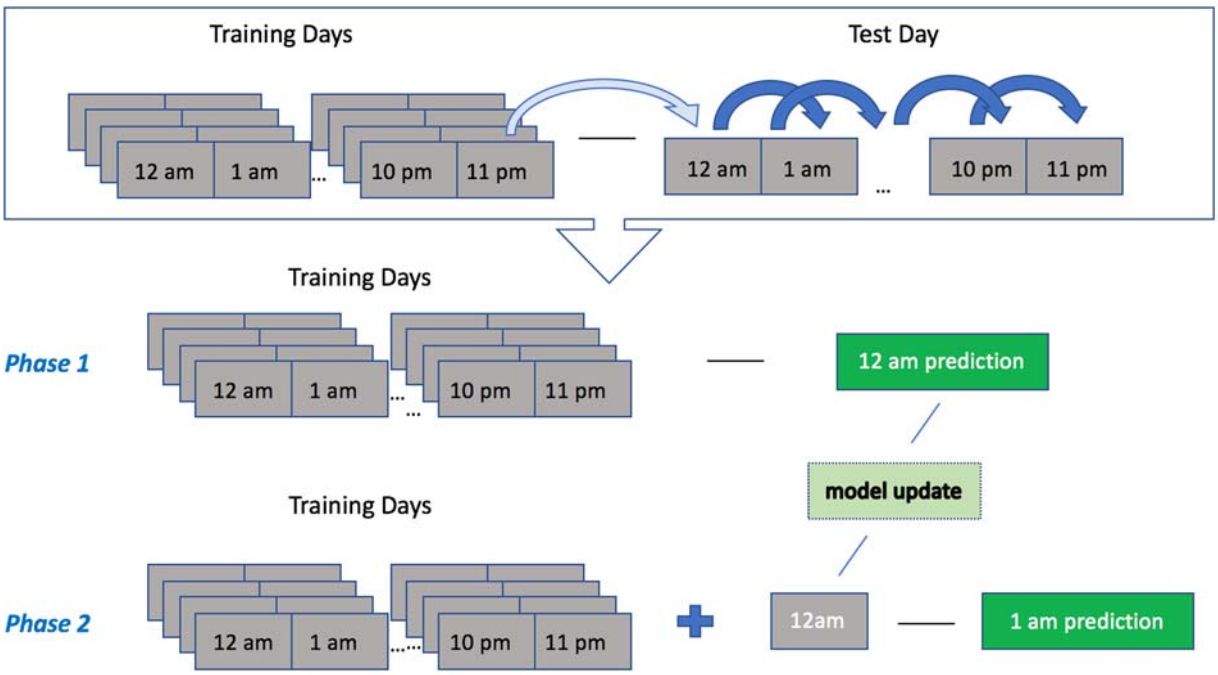

Fig. 3. Scenario 2 - Multi-step predictions with model update

References

Krygsman, S., Dijst, M. and Arentze, T., 2004. Multimodal public transport: an analysis of travel time elements and the interconnectivity ratio. Transport Policy, 11(3), pp.265-275.

Moreira-Matias, L., Gama, J., Ferreira, M., Mendes-Moreira, J. and Damas, L., 2013. Predicting taxi–passenger demand using streaming data. IEEE Transactions on Intelligent Transportation Systems, 14(3), pp.1393-1402.

Van Oort, N., Brands, T. and de Romph, E., 2015. Short term ridership prediction in public transport by processing smart card data. Transportation Research Record, (2015).

Pereira, F.C., Rodrigues, F. and Ben-Akiva, M., 2015. Using data from the web to predict public transport arrivals under special events scenarios. Journal of Intelligent Transportation Systems, 19(3), pp.273-288.

Chan, J.W., Chang, V.L., Lau, W.K., Law, L.K. and Lei, C.J., 2016. Taxi App Market Analysis in Hong Kong. Journal of Economics, Business and Management, 4(3).

Yuan, J., Zheng, Y., Zhang, L., Xie, X. and Sun, G., 2011, September. Where to find my next passenger. In Proceedings of the 13th international conference on Ubiquitous computing (pp. 109-118). ACM.

Xu, J., Rahmatizadeh, R., Bölöni, L. and Turgut, D., 2017. Real-Time Prediction of Taxi Demand Using Recurrent Neural Networks. IEEE Transactions on Intelligent Transportation Systems.

Davis, N., Raina, G. and Jagannathan, K., 2016, November. A multi-level clustering approach for forecasting taxi travel demand. In Intelligent Transportation Systems (ITSC), 2016 IEEE 19th International Conference on (pp. 223-228). IEEE.

Tang, X., Yang, C. and Zhou, J., 2009, September. Stock price forecasting by combining news mining and time series analysis. In Web Intelligence and Intelligent Agent Technologies, 2009. WI-IAT'09. IEEE/WIC/ACM International Joint Conferences on (Vol. 1, pp. 279-282). IEEE.

Si, J., Mukherjee, A., Liu, B., Li, Q., Li, H. and Deng, X., 2013. Exploiting Topic based Twitter Sentiment for Stock Prediction. ACL (2), 2013, pp.24-29.

O'Connor, B., Balasubramanyan, R., Routledge, B.R. and Smith, N.A., 2010. From tweets to polls: Linking text sentiment to public opinion time series. ICWSM, 11(122-129), pp.1-2.

TLC - New York City Taxi & Limousine Commission, Taxi and limousine commission trip record data, Available: http://www.nyc.gov/html/tlc/html/about/trip record data.shtml (2017).

Best concert venues in NYC, https://www.nyc.com/nyc-guides/best_ concert_venues_in_nyc.308/, accessed: 2017-08-14.