REAL-TIME TAXI DEMAND PREDICTION USING TRANSFER LEARNING

Ioulia Markou[1], Filipe Rodrigues[1] and Francisco C. Pereira[1]

[1]DTU Management Engineering – Transport Modelling

Introduction

In urban systems, nature, economy, environment, and many other settings, there are multiple simultaneous phenomena happening that are of interest to model and predict. Phenomena of relevance include travel times and travel demand from different transport modes for different areas in a city, infrastructure conditions, weather impacts, parking availability, energy demand, emissions, and so on. For practical and historical reasons, the approach has been to focus on each phenomenon separately, by gathering data, designing and estimating one model with one problem in mind. However, because these scenarios are frequently rich in correlations between those phenomena, knowing the value of one response variable may contribute to improving prediction quality of other response variables.

Demand prediction is one of the non-trivial research subjects that attracts particular interest due to its inherent complexity. Taxi demand is a characteristic example of a challenging research problem, because of the many parameters of underlying information. A taxi differs from other modes of public transport where the pick-up and drop-off locations are determined by the service provider, not by the passenger. Taxi calling platforms, such as Uber, Grab and Beat are becoming increasingly popular, especially in situations of traffic congestion, 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. This trend, therefore, 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.

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, Moreira-Matias et al., 2013). A unified linear regression model that outperforms other popular non-linear models in the prediction accuracy of the Unit Original Taxi Demand (UOTD) is proposed by Tong et al. (2017). A simple model structure that eliminates the need for repeated model redesign proves to be able to behave better in prediction scenarios with high-dimensional features.

All these research studies follow the typical approach of a demand prediction model formulation that explores information related to the area of interest. The independent models for different areas of interest follow the converse direction, where each response variable is separately modelled, with its own dataset. They have the benefit of scalability and flexibility (e.g. different models can follow radically different function forms), but they ignore correlations between different response variables.

In this study we explore the utility of information from other areas for a selected area taxi demand prediction optimization. This collaboration is defined through a correlation structure that is strongly dependent on domain, and itself potentially dynamic. The time window of the study comprehends 4 years (2013-2016) and our work is focused on New York City (NYC) using a large-scale public dataset of 1.1 billion taxi trips.

Data description and preparation

In this research we work with a taxi dataset 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 decided to focus our study on the area of Manhattan and more specifically on the five neighborhoods shown in Figure 1. The selected areas show significant demand fluctuations within a day, due to many entertainment options that they offer to tourists and residents.

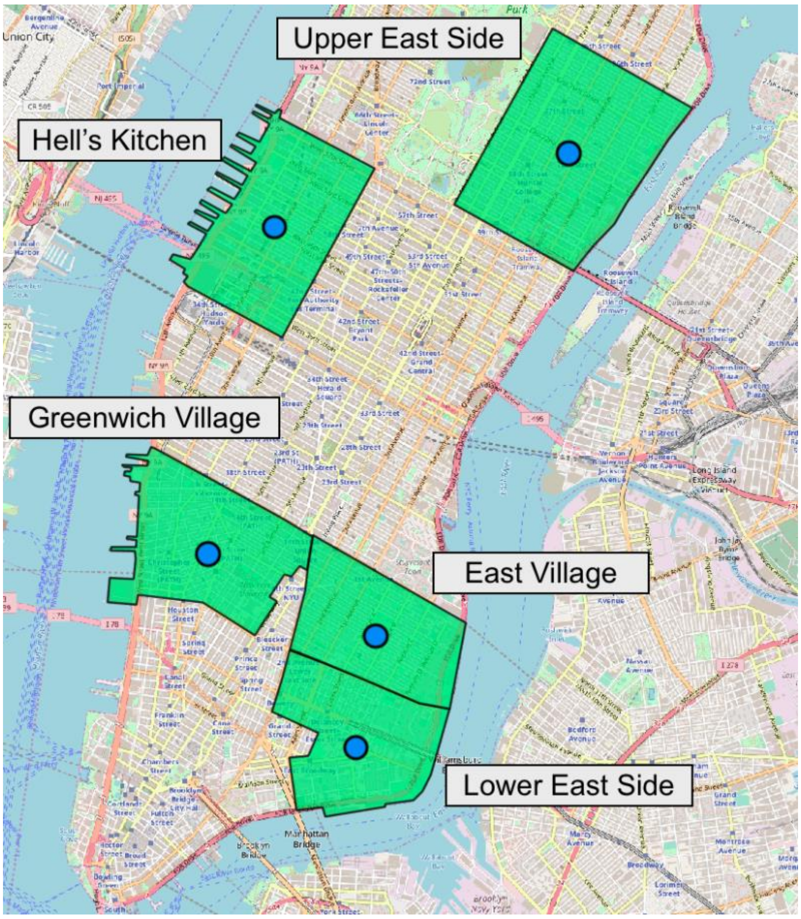

Figure 1 Selected areas in Manhattan

Methodology

Our model aims to predict the number of taxi pickups from a given area, occurring at a given hour interval of a given day by taking into consideration short-term time series trends. In other words, the proposed approach is focused on hourly short-term predictions.

The raw dataset that we obtained from TLC includes fields capturing pick-up and drop-off dates/times, pickup and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts. The first three years of our dataset (2013-2015) is our model training set and the first six months of 2016 (January 2016 - June 2016) our test set. For the baseline model of each area the corresponding pickups and drop-offs aggregated by hour are used:

$$\widehat{Y_{t+1}} = \widehat{\beta_0} + \widehat{\beta_{1,1}} Y_t + \cdots + \widehat{\beta_{1,6}} Y_{t-6} + \widehat{\beta_{2,1}} D_t + \cdots + \widehat{\beta_{2,6}} D_{t-6}$$

where Y represents the taxi pickups counts, D represents the hourly drop-off counts, and $\widehat{\beta_0}, \widehat{\beta_{1,1}}, \ldots, \widehat{\beta_{2,12}}$ are estimated using data through period t. It is an 6th Order Autoregressive Model, namely Y_t is regressed against $Y_{t-1}, Y_{t-2}, \ldots, Y_{t-6}$.

It is obvious that, if we can somehow transfer relevant knowledge from one model to the others, we would improve them, even with very minimal new data available. Therefore, for the enhanced model of each area the correlation of pickups and drop-offs of the corresponding area with the pickups and drop-offs of the other four areas is examined using again the training set. For the correlation matrix calculation and the final model formulation, only the pickup and drop-off lags are used, since we don't want to introduce any information about the demand at the timestamp *t*, which is our target time interval. Figure 2 gives an overview of the implemented methodology. The main square represents the correlation matrix of the five areas. The red color gradations indicate that the Pearson correlation coefficient is above 0.5, while the blue color gradations below 0.5.

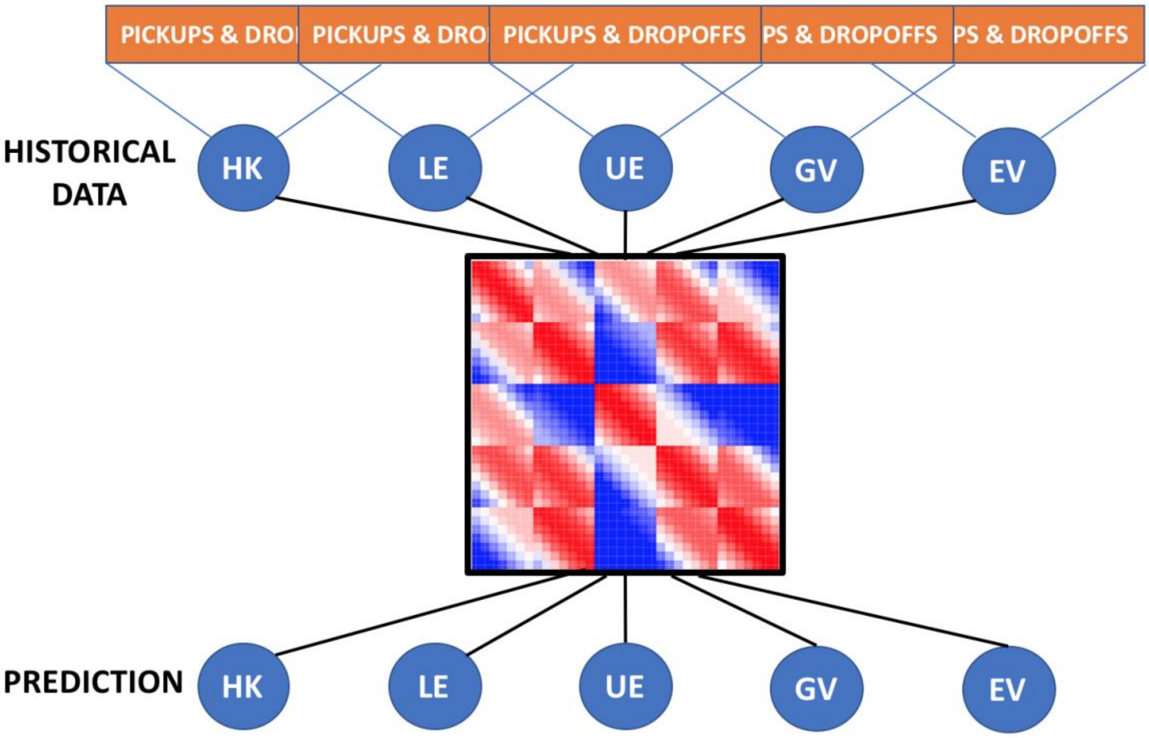

Figure 2 Implemented methodology

We chose to explore the proposed methodology using Linear Regression (LR) for its simplicity and interpretability and Gaussian processes (GP) because they are flexible enough to represent a wide variety of interesting model structures. The proposed models were implemented with the scikit-learn machine learning library in Python.

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), computed as follows:

$$MAE(\hat{x}) = \frac{1}{n}\sum_{t=1}^{n}|x_i - \hat{x}_i|$$

$$RMSE(\hat{x}) = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(x_i - \hat{x}_i)^2}$$

$$R^2 = \frac{\sum_{t=1}^{n}(\hat{x}_i - \bar{x})^2}{\sum_{t=1}^{n}(x_i - \bar{x})^2}$$

where n denotes the number of instances in the dataset, $\hat{x}_i$ is the predicted taxi pick-ups count for the i_{th} instance, x_i is the corresponding true pick-up count and $\bar{x}$ is the mean of the observed counts.

Results

The parameters of other areas that showed high correlation (ϱ>0.85) with the pickups and drop-offs of the examined area are included into our enhanced model. The analysis results are summarized in Table 1. Our baseline model with pickups and drop-offs lags of the studied area makes satisfactory predictions. But the enhanced models of each area are even more accurate. For example, Hell's Kitchen enhanced model take into consideration the high correlation of its pickups with the corresponding value of Greenwich Village's neighborhood (ϱ>0.86). High correlation exists also for the drop-offs of those areas (ϱ>0.91). It was therefore considered useful to include them in the model.

Table 1 Summary of results

	LR - BASELINE MODEL			LR - ENHANCED MODEL		
	R2	RMSE	MAE	R2	RMSE	MAE
HELL'S KITCHEN	0.867	63.616	48.166	0.881	60.109	45.419
LOWER EAST SIDE	0.951	34.281	20.543	0.964	29.562	17.577
UPPER EAST SIDE	0.954	133.883	96.642	0.959	126.111	92.945
GREENWICH VILLAGE	0.946	60.433	43.161	0.952	56.671	40.863
EAST VILLAGE	0.950	48.429	31.705	0.957	45.292	30.247

	GP - BASELINE MODEL			GP - ENHANCED MODEL		
	R2	RMSE	MAE	R2	RMSE	MAE
HELL'S KITCHEN	0.896	56.287	40.711	0.912	51.626	36.588
LOWER EAST SIDE	0.957	32.353	16.476	0.959	31.528	15.796
UPPER EAST SIDE	0.979	90.115	63.707	0.981	86.768	61.457
GREENWICH VILLAGE	0.961	51.040	34.037	0.967	47.481	33.407
EAST VILLAGE	0.953	47.214	26.849	0.958	44.703	25.067

A characteristic example of the small but useful optimization of the prediction results is shown in Figure 3. The black line corresponds to the true demand values, while the green line to the enhanced model prediction results. It is clear, that the latter is better able to capture the demand peak around 2am, and to avoid the demand overestimation during the evening (20:00-23:00).

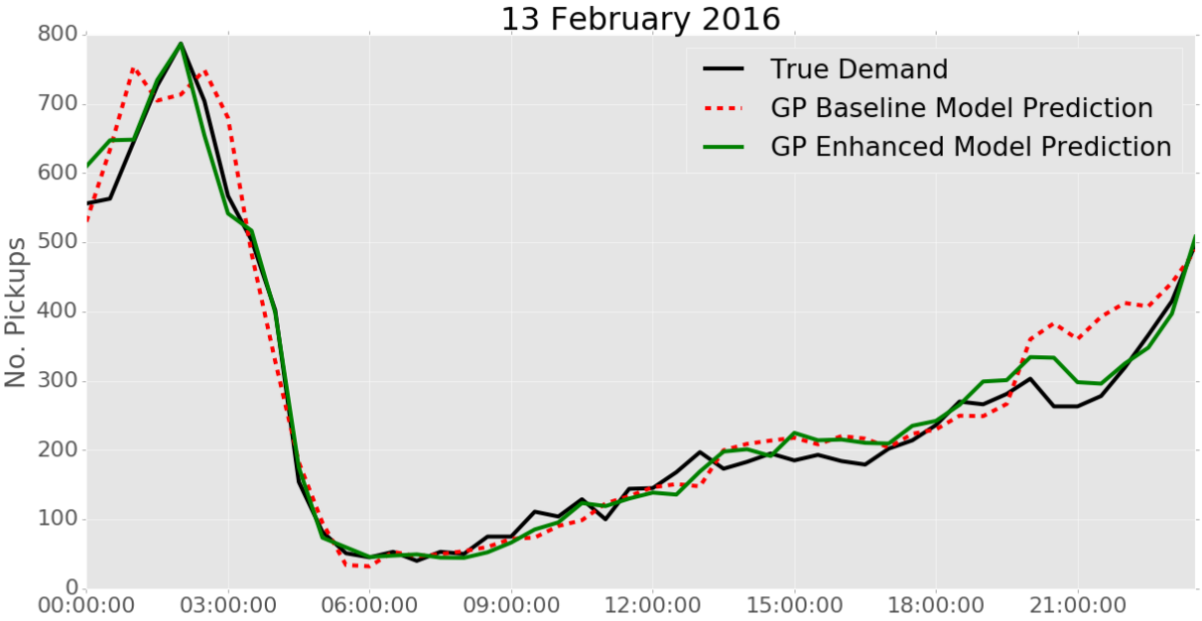

Figure 3 Model prediction results using GPs

References

- J. Yuan, Y. Zheng, L. Zhang, X. Xie, and G. Sun, "Where to find my next passenger," in Proceedings of the 13th international conference on Ubiquitous computing. ACM, 2011, pp. 109–118.
- J. Xu, R. Rahmatizadeh, L. Boloni, and D. Turgut, "Real-time prediction of taxi demand using recurrent neural networks," IEEE Transactions on Intelligent Transportation Systems, 2017.
- N. Davis, G. Raina, and K. Jagannathan, "A multi-level clustering approach for forecasting taxi travel demand," in Intelligent Transportation Systems (ITSC), 2016 IEEE 19th International Conference on. IEEE, 2016, pp. 223–228.
- 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.
- Y. Tong, Y. Chen, Z. Zhou, L. Chen, J. Wang, Q. Yang, J. Ye, and W. Lv, "The simpler the better: a unified approach to predicting original taxi demands based on large-scale online platforms," in Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2017, pp. 1653–1662.
- "TLC Trip Record Data", http://www.nyc.gov/html/tlc/ html/about/trip record data.shtml, 2017, [Online; accessed 21-November-2017].