Short-term bike-sharing demand forecasting incorporating multiple sources of uncertainties

Jingyi Cheng[a], Gonçalo Correia[a], Oded Cats[a] and Shadi Sharif Azadeh[a]

[a] Transport and Planning, Delft University of Technology, Delft, The Netherlands
j.cheng-1@tudelft.nl, g.correia@tudelft.nl, o.cats@tudelft.nl, s.sharifazadeh@tudelft.nl

Extended abstract submitted for presentation at the Conference in Emerging Technologies in Transportation Systems (TRC-30)
September 02-03, 2024, Crete, Greece

April 30, 2024

Keywords: Short-term demand forecasting, Shared mobility, Machine learning

1 Introduction and Problem Description

Demand uncertainty poses crucial challenges to the operations of bike-sharing services, including real-time supply-demand re-balancing via fleet allocation and dynamic pricing. Recent research has explored the use of machine learning techniques to capture the complex interdependencies between short-term demand for bike-sharing and various influencing factors, including endogenous elements that shape demand patterns and exogenous environmental factors that cause demand fluctuations. Li *et al.* (2023) identify semantic neighbors of bike-sharing service points and introduce a hybrid CNN and LSTM forecasting model that utilizes neighbors' demand series to capture the spatial dependencies. Liang *et al.* (2022) propose a multi-modal demand prediction framework that integrates NYC subway usage data to enhance the forecasting of nearby ride-hailing demand, and vice versa. Lv *et al.* (2021) introduce a framework that firstly generates categorical features to describe the demand pattern of subway stations, which is then used for the predictions of the drop-off demand of bike-sharing within the neighborhood of these stations. Although predictions have been improved by using multi-source data, the potential benefits of modeling the demand time series by separately recognizing two types of influences, the consistent patterns and local fluctuations, remain largely unexplored. Secondly, the potential of integrating passenger flow fluctuations from interconnected public transport systems into short-term demand forecasting for shared mobility services remains under-explored in the existing literature. Conducting case studies can offer richer insights into the advantages of incorporating cross-modal factors into forecasting algorithms and inspire policy designs for multimodal coordinating services. Lastly, the operational efficiency of bike-sharing services could significantly benefit from enhancing prediction granularity to every 15 minutes at each service point or grid level. While most previous studies focus on short-term predictions of an hour or more, the state-of-the-art models they use may struggle with zero-inflated demand time series resulting from higher forecasting granularities.

To address these gaps in the literature on demand forecasting for shared mobility, we propose a per 15-minute demand forecasting framework that tackles the predictive challenges arising from complex demand patterns and the demand uncertainties associated with various factors, including seasonality, weather and the varying passenger flows from interconnected metro stations. In addition, we introduce a hurdle approach that sequentially predicts regular demand patterns and temporal demand fluctuations.

Our study utilizes the historical trip data from 2022/10/01 to 2022/12/31 of Washington

D.C. Capital bike-sharing service[1]. On average, this dock-based bike-sharing service attracts more than 6000 pickups per day from 330 service points across the city. The original data is processed through an aggregation of pickups per 15-minute intervals for each service point, with 96 time windows each day. Preliminary data analysis indicates that the bike-sharing demand in Washington D.C. exhibits both complex temporal seasonality patterns and spatial patterns. The average daily demand is found relatively consistent across different weekdays. Additionally, the hourly demand pattern exhibits dual peaks between 7:00-9:00 AM and 4:00-7:00 PM on weekdays. Nevertheless, the weekly-hourly demand exhibits spatially distinct patterns across various service points in the city, with notable differences observed between downtown and suburban areas, as well as between metro-connected and not connected service points. Weather conditions are found affecting the total number and spatial distribution of demand in our data.

2 Multivariate Short-term Demand Forecasting Framework

We leverage four types of inputs to predict the bike-sharing demand for the next 15-minute intervals, covering the influence of seasonality, real-time weather conditions, passenger flow from interconnected metro services, and other underlying events. *hour-of-a-day* and *day-of-a-week* are adopted as seasonal features. The hourly precipitation (millimeters per hour), wind speed (meter per second), and temperature (Celsius) as included as weather features. To capture the temporal effects on demand, we also consider the previous demand observations of the past M 15-minute time windows, denoted by $y_{t-j:t-j+1}$. Following the suggestion from Ma *et al.* (2018), we consider a pair of bike-sharing service point and metro station to be interconnected if their distance is within 300 meters, resulting in 92 pairs of metro-connected service points. We adopt the number of passenger check-outs during time window t for station s, c_t^s, as a feature that measures the passenger flow from public transit, which is aggregated from historical passenger movement data from local public transit operator WMATA.

Short-term demand predictors for bike-sharing should handle multiple sources of input information while capturing the non-linearity dependency between these features and demand, and being computationally fast for prediction generation and storage efficient to apply in practice. Meeting these criteria, XGBoost and LSTM have also achieved state-of-the-art performance in previous demand forecasting studies (Xu *et al.*, 2018, Li *et al.*, 2023). Trained using a gradient boosting method, XGBoost models generate predictions by combining the outcomes from a sequence of regression trees. LSTM models describe the complex temporal pattern by capturing the dynamics of both long- and short-term dependencies on the time series. Since the bike-sharing demand time series is jointly influenced by both contextual and temporal effects, we utilize both approaches as the basic models in this study.

In preliminary data analysis, we observed that bike-sharing demand is predominantly driven by reoccurring demand patterns described by seasonality and weather. However, we also identified periods of significant demand deviations that exhibit local sequential patterns. Hence, we introduce a *two-stage hurdle framework* to firstly estimate the expected demand given contextual inputs, then fine-tune the predictions by considering the short-term temporal effects. The final demand prediction $\hat{y}_{i,t}$ is therefore produced as a nonlinear combination of regular demand estimation $\hat{y^d_{i,t}}$ and the recent observations $y_{i,t-1}, \cdots, y_{i,t-M}$,

$$\hat{y}_{i,t} = f(\hat{y}^1_{i,t}, y_{i,t-1}, \cdots, y_{i,t-M}), \tag{1}$$

, where $\hat{y}^1_{i,t}$ is estimated by service point-wise predictor $g^1(\cdot)$ with seasonal, weather (and passenger flow features). We can train both first- and second-stage predictors $f(\cdot)$ and $g^1(\cdot)$ for each service point individually. Alternatively, we can train $f(\cdot)$ as a general second-stage predictor applied for all service points, if we assume service point-wise differences are fully described by

[1] This data is available for download on Capital's open-sourced data portal `https://capitalbikeshare.com/system-data`

the predictor $g^1(\cdot)$. In our experiment, XGBoost is adopted as the first stage model for regular demand estimation, because it effectively captures complex dependencies among contextual features, while filtering out anomalies and randomness in the demand time series. Both LSTM and XGBoost can be employed as the second-stage model to capture the temporal demand deviation dynamics. Moreover, compared to the 'one-stage' models where all features are handled simultaneously, hurdle models improve model training efficiency and parameter estimation quality through a reduction of dimensionality by splitting the handling of inputs into two stages. Based on a partial auto-correlation (PAC) analysis over the demand time series, we choose the past $M = 6$ previous demand as temporal inputs.

Table 1 – *Overview of short-term forecasting models and their corresponding input features.*

Model Category	**Model Name**	**Features**			
		Seasonality	**Weather**	**Prev. Obs.**	**Metro**
Partial Inputs	*Historical Average*	✓	✓		
	Univariate LSTM			✓	
	Contextual XGBoost	✓	✓		
	Contextual XGBoost with PT	✓	✓		✓
One-stage	*Regular XGBoost with PT*	✓	✓	✓	✓
	Regular LSTM with PT	✓	✓	✓	✓
Hurdle	*General Hurdle XGBoost with PT*	✓	✓	✓	✓
	Individual Hurdle XGBoost with PT	✓	✓	✓	✓
	General Hurdle XGBoost-LSTM with PT	✓	✓	✓	✓
	Individual Hurdle XGBoost-LSTM with PT	✓	✓	✓	✓

In this study, we design and evaluate three categories of models with different inputs, summarized in Table 1. Although we only consider the metro passenger flow as a public transit (PT) feature, our forecasting framework can be extended to other types of connected PT flows as inputs. However, if a bike-sharing service point is not connected to any metro stations, its demand predictor simply excludes this feature and remains the same as the cases where PT information is not considered. We analyze the contribution of each source of input by comparing the forecasting performance of models where only a subset of features is used. Specifically, we include the *Historical Average* as a benchmark that calculates the average demand based on seasonal features. 'General' and 'Individual' denote whether the second-stage predictor is generally applied for all service points, or specifically trained per individual service point.

3 Preliminary Results and Discussion

The historical trip data, covering a period of 92 days, of the bike-sharing service, is separated into a training set and a testing set. We allocate the initial 69 days of data, equivalent to 75% of the total, for model training. The last 23-day data is reserved for model testing. The hyper-parameter tuning for all XGBoost models is done via a Bayesian optimization-based approach, combined with five-fold cross-validation on the training data. We adopt a Poisson loss function with regularization for the training of XGBoost models.

Table 2 – *Per 15-minute demand forecasting performance of various models.*

Model Category	**Model Name**	**Testing data**		**Training data**	
		MAE	RMSE	MAE	RMSE
Partial Inputs	*Historical Average*	0.2495	0.4616	0.2785	0.5559
	Univariate LSTM	0.2498	0.4689	0.3735	0.6956
	Contextual XGBoost	0.1916	0.4185	0.2535	0.5068
	Contextual XGBoost with PT	0.1897	0.4180	0.2468	0.4925
One-stage	*Regular XGBoost with PT*	0.1914	**0.4150**	0.2348	0.4564
Hurdle	*General Hurdle XGBoost with PT*	0.2181	0.4410	0.3430	0.6509
	Individual Hurdle XGBoost with PT	**0.1890**	0.4480	**0.2104**	**0.4327**

The overall prediction accuracy for various models on both testing and training data is reported in Table 2 with metrics Mean absolute errors (MAE) and root mean squared errors (RMSE). Our study is a work in progress. We are still working on the LSTM models and will

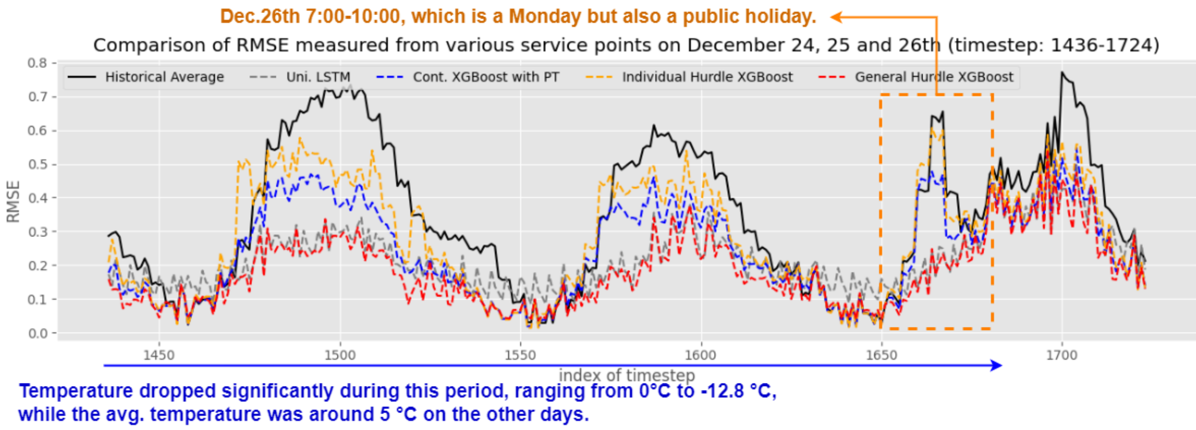

Figure 1 – *Comparison of the averaged RMSEs measured for each time window during a three-day demand abnormal period within testing data, from 2022/12/24 to 2022/12/26.*

finish the related experiments before the conference. Among the models reported in Table 2, the least prediction errors are obtained where all input features are applied. The performance improvement over models with partial inputs suggests that jointly including all four sources of inputs enables a better capture of demand dynamics. Specifically, for bike-sharing service points interconnected to metro services, the RMSE is reduced by 11.0% and 34.6% for predictions over the testing and training data respectively, by including passenger flow features in contextual XGBoost models.

In the testing dataset, we observed an abnormal period from December 24 to December 26, during a national holiday coinciding with temperature rapidly dropping to minus 10 degrees. Interestingly, although the lowest MAE over testing data and lowest MAE and RMSE over training data have been achieved by *Individual Hurdle XGBoost with PT*, Figure 1 suggests *General Hurdle XGBoost with PT* to work better during this abnormal scenario. As circled out in Figure 1, the predictions by the general hurdle approach manages to adapt to the sequential temporal effects. It indicates that the combined view of deviation series from other service points during model training improves forecasting quality during abnormal periods with large demand deviations, which may provide crucial insights for operators. Therefore, as the next step of this study, we seek to refine the hurdle method to incorporate a sophisticated spatial connecting structure, possibly via CNNs (Li *et al.*, 2023) or GCNs (Lee & Rhee, 2022).

References

Lee, Kyungeun, & Rhee, Wonjong. 2022. DDP-GCN: Multi-graph convolutional network for spatiotemporal traffic forecasting. *Transportation Research Part C: Emerging Technologies*, **134**, 103466.

Li, Xinyu, Xu, Yang, Zhang, Xiaohu, Shi, Wenzhong, Yue, Yang, & Li, Qingquan. 2023. Improving short-term bike sharing demand forecast through an irregular convolutional neural network. *Transportation research part C: emerging technologies*, **147**, 103984.

Liang, Yuebing, Huang, Guan, & Zhao, Zhan. 2022. Joint demand prediction for multimodal systems: A multi-task multi-relational spatiotemporal graph neural network approach. *Transportation research part C: emerging technologies*, **140**, 103731.

Lv, Ying, Zhi, Danyue, Sun, Huijun, & Qi, Geqi. 2021. Mobility pattern recognition based prediction for the subway station related bike-sharing trips. *Transportation Research Part C: Emerging Technologies*, **133**, 103404.

Ma, Xinwei, Ji, Yanjie, Yang, Mingyuan, Jin, Yuchuan, & Tan, Xu. 2018. Understanding bikeshare mode as a feeder to metro by isolating metro-bikeshare transfers from smart card data. *Transport policy*, **71**, 57–69.

Xu, Chengcheng, Ji, Junyi, & Liu, Pan. 2018. The station-free sharing bike demand forecasting with a deep learning approach and large-scale datasets. *Transportation research part C: emerging technologies*, **95**, 47–60.