A Neighborhood-augmented LSTM Model for Taxi-Passenger Demand Prediction

Tai Le Quy*, Wolfgang Nejdl*, Myra Spiliopoulou**, Eirini Ntoutsi*

*Leibniz University Hannover & L3S Research Center, LUH, Hannover

**Otto-von-Guericke University, Magdeburg

Würzburg, 16.09.2019

Content

- Introduction
- Methodology
- Dataset
- Experiments
- Results
- Conclusion and outlook

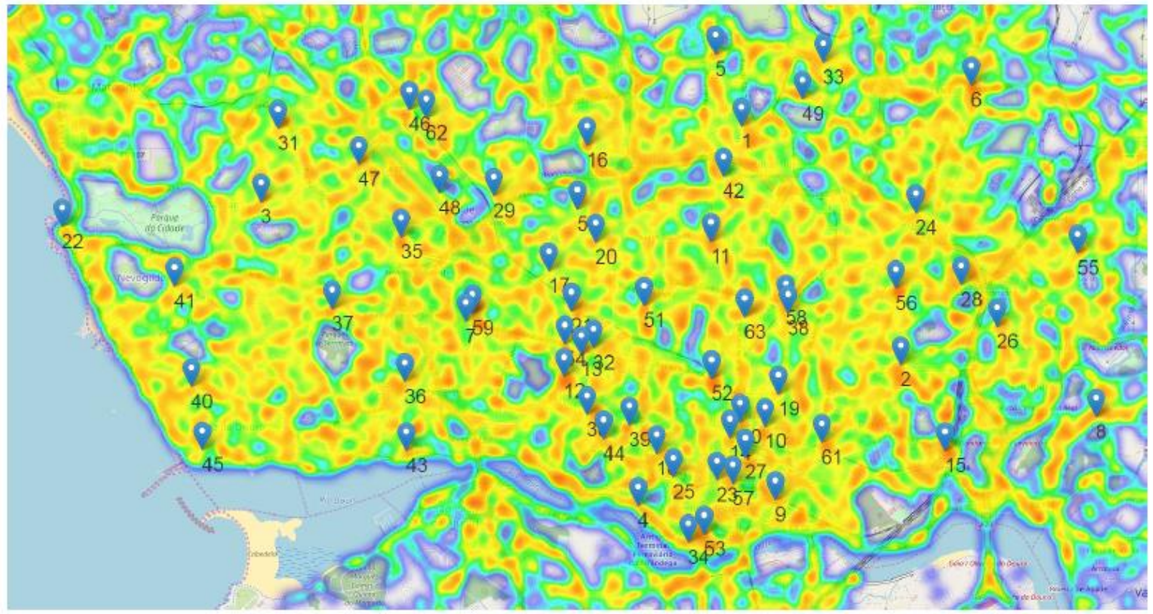
Image source: http://www.pngall.com/taxi-cab-png/download/12771

Introduction

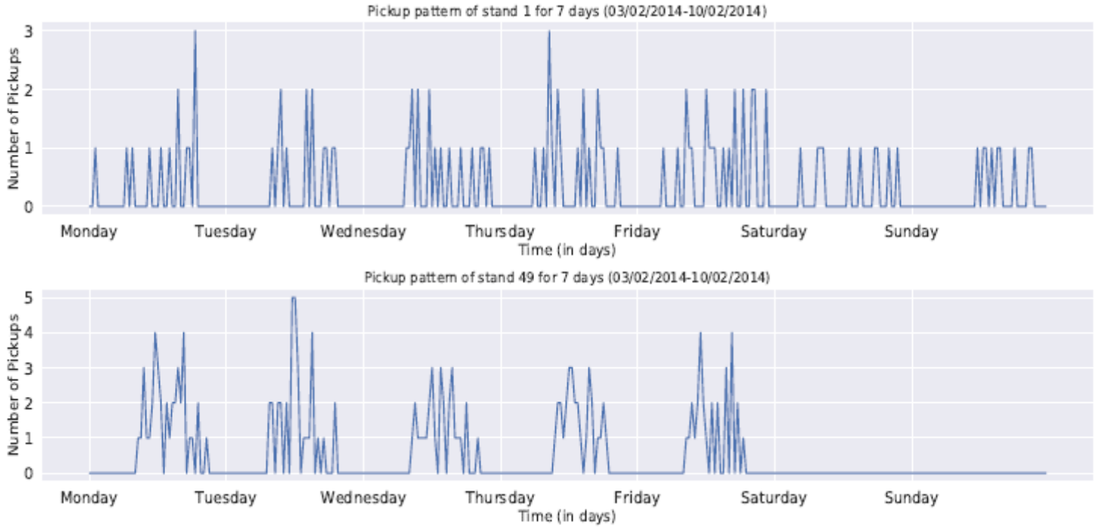
- Taxi companies
  - Improve the levels of passenger satisfaction and maximal profit
  - Balance the relationship between the passenger demand and the number of running taxi vehicles
- Taxi-Passenger Demand Prediction
  - It’s useful for drivers in making decision moving to pick up passengers in a particular region in the city
  - Spatial information is useful for prediction task

Introduction

- Motivation
  - The intuition: nearby taxi-stands might have similar demands

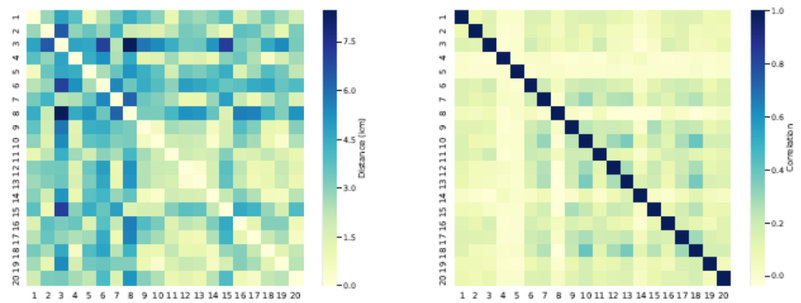


Spatial distribution of the taxi-stands
Numbers 1- 63 indicate the IDs of the stands



Pickup demand history for nearby taxi-stands 1 and 49.

- Goal
  - Develop a neighborhood-augmented LSTM model to predict the taxi-passenger demand



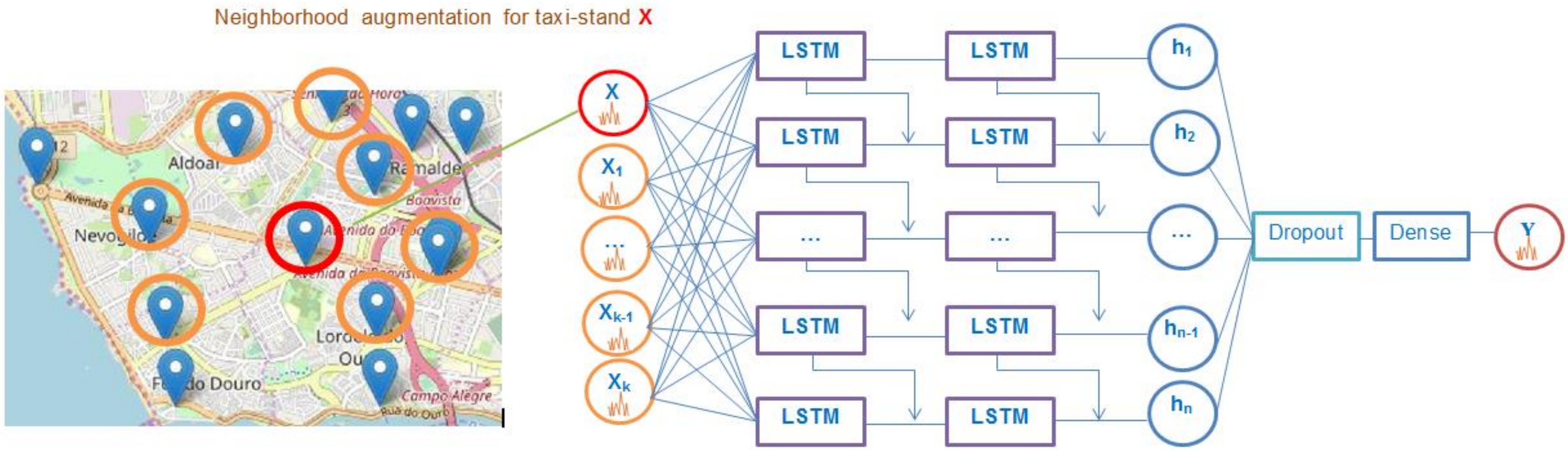
Spatial proximity (left) vs pickup demand correlation (right) between taxi-stands (based on dataset D1).

Problem denition

- Let $S = \{s_1; s_2; ..; s_N\}$ be the set of predened N taxi-stands in a city
- $X_s = \{X_{s;0}; X_{s;1}; ..; X_{s;t}\}$ to be a discrete time series modeling the taxi-demand for stand s
  - based on an aggregation period of P-minutes
- Our goal is to build a model which predicts the demand $X_{s;t+1}$ for the next time point *t + 1* at taxi-stand *s*.

Methodology

- Neighborhood-augmented LSTM



The architecture of the neighborhood-augmented LSTM.

Neighborhood-augmented LSTM

- Algorithm
  - An LSTM model for each taxi-stand
  - Input of LSTM: (*k+1*) dimensional vector of taxi-stand *s* and its *k*-neighbors
  - Output: taxi demand for taxi-stand *s*

input : Taxi demand dataset; k-number of neighbors
output: Prediction model M_s for taxi-stand s

1 *//Data augmentation*
2 X_s: the demand history of taxi-stand s up to time t
3 $X'_s \leftarrow X_s$ *//extended representation*
4 $\{Neighbors_s\}$: the set of k nearest taxi-stands to s
5 **for** $i \leftarrow 1$ **to** $|\{Neighbors_s\}|$ **do**
6 X_i: the demand history of taxi-stand i
7 $X'_s \leftarrow Extend(X'_s, X_i)$
8 **end**
9 Normalize features
10 *//Train on the augmented data*
11 $M_s \leftarrow LSTM(X'_s)$

Algorithm 1: Neighborhood-augmented LSTM model training

Neighborhood-augmented LSTM

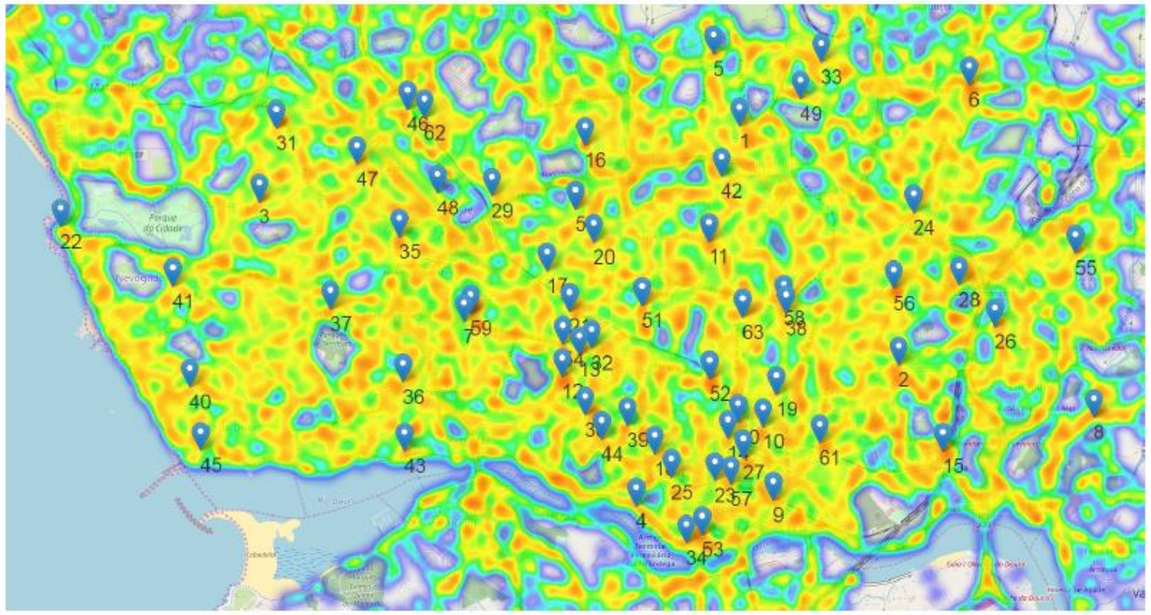
1. Input (X'_s, the extended description of stand s; look back value = 5 (see Section 4.4.)
2. LSTM (N=200, optimizer = 'Adamax', Activation function = 'tanh', loss= 'mean squared error', batch size = 100 (see Section 4.4.))
3. Full connected LSTM(N=200, Activation function ='tanh')
4. Dropout =0.7 (see Section 4.4.)
5. Dense (Activation function = 'tanh')

Dataset

- Porto city in Portugal
  - Period: July 2013 to June 2014
  - Records: 1.710.670
  - 9 features
  - 63 taxi-stands
- Two versions of dataset for experiment
  - D1: all trips departing from taxi-stands (817.861 instances)
  - D2: all trips (1.706.572 instances).
    - Assign trips (do not start from a taxi-stand) to their closest taxi-stand based on distance.

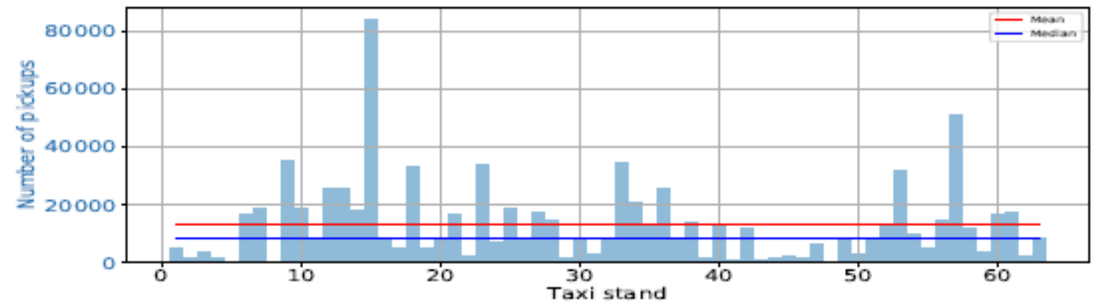
Dataset Characteristics

- Spatial distribution

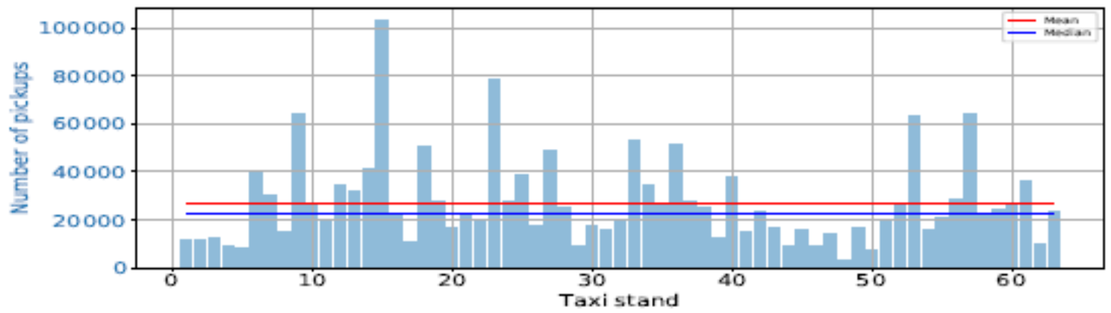


Spatial distribution of the taxi-stands
Numbers 1- 63 indicate the IDs of the stands

Pickup distribution



Pickup distribution per taxi-stand on D1



Pickup distribution per taxi-stand on D2

Experiment

- Experimental setup
  - Set aggregation period at 30 minutes
  - 70% data for training, 30% data for testing (by the time series)
  - Validation, turning parameters on data of taxi stand 1
- Baselines
  - Simple Moving Average
  - Linear Regression
  - Random Forest Regression
  - XGBoost Regression

- Evaluation measure
  - symmetric MeanAbsolute Percentage Error (sMAPE)

$$sMAPE_s = \frac{100\%}{t} \sum_{i=1}^{t} \frac{| X_{s,i} - \hat{X}_{s,i} |}{| X_{s,i} | + | \hat{X}_{s,i} | + c}$$

in which, X_s and $\hat{X}_s$ are the true and predicted demand, c=1

  - Mean Squared Error (MSE)

Results

- Taxi-demand prediction quality results
  - Neighborhood-augmented LSTM outperforms other models in term of sMAPE.

Model	Training		Testing	
	sMAPE (%)	MSE	sMAPE (%)	MSE
Simple Moving Average			23.34	1.721
Linear Regression	24.37	1.61	24.52	1.765
Random Forest Regression	**16.83**	**0.383**	24.25	1.660
XGBoost Regression	23.90	1.391	23.91	**1.585**
LSTM	18.37	1.659	18.54	1.839
Neighborhood-augmented LSTM	17.32	1.465	**17.63**	1.682

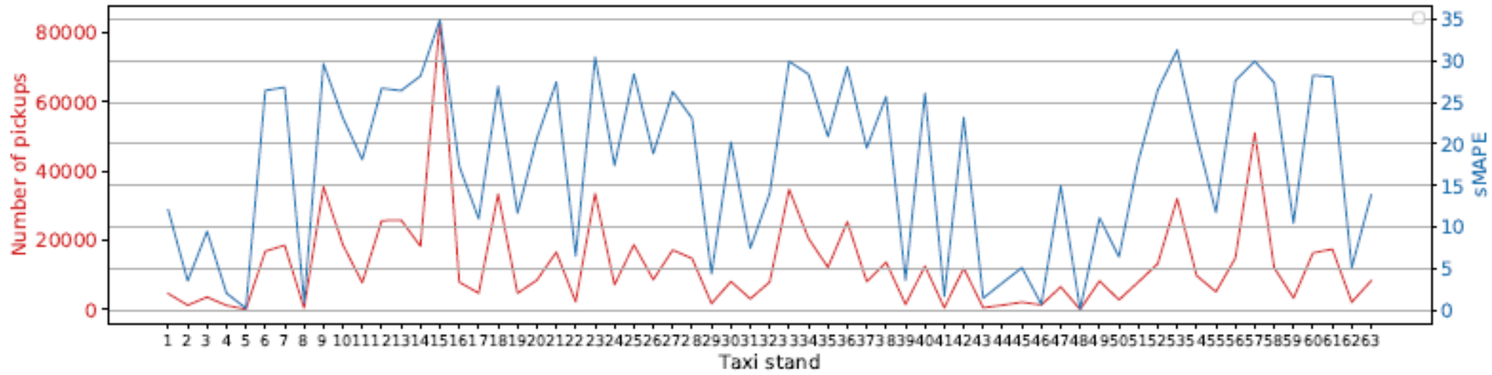
Prediction quality of the different models on D1
K=15

Model	Training		Testing	
	sMAPE (%)	MSE	sMAPE (%)	MSE
Simple Moving Average			30.33	**5.369**
Linear Regression	30.78	4.206	31.23	5.988
Random Forest Regression	**18.49**	**0.715**	31.03	5.503
XGBoost Regression	30.466	3.605	30.51	5.449
LSTM	27.03	4.16	27.22	6.660
Neighborhood-augmented LSTM	25.88	3.84	**26.07**	6.444

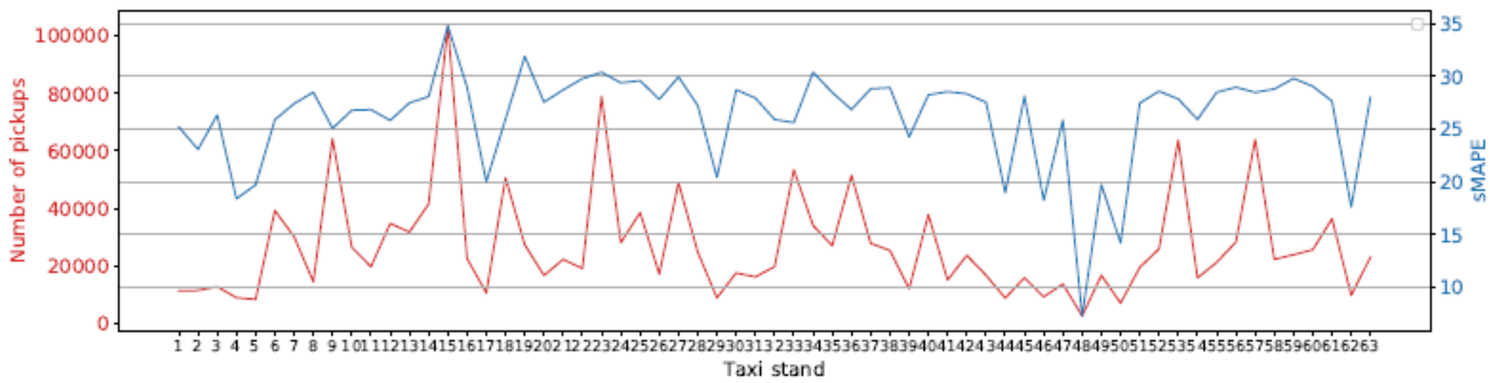
Prediction quality of the different models on D2
K=25

Results

- Performance of models on taxi-stand



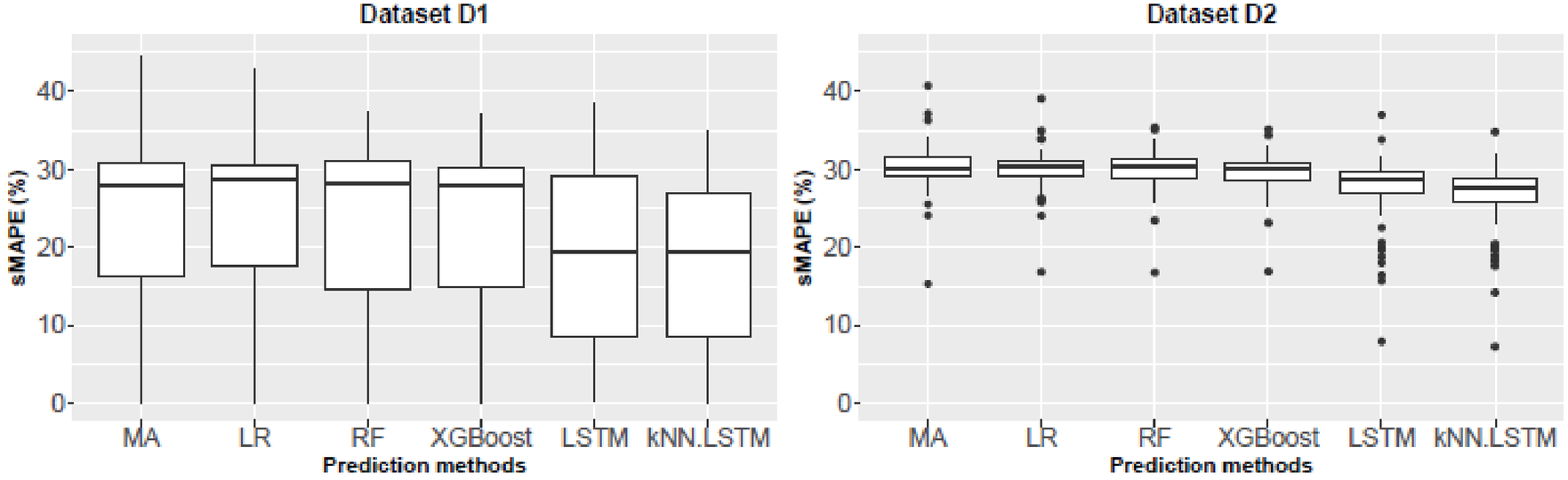
Real demand vs neighborhood-augmented LSTM error
across dierent taxi-stands for dataset D1



Real demand vs neighborhood-augmented LSTM error
across dierent taxi-stands for dataset D2

Results

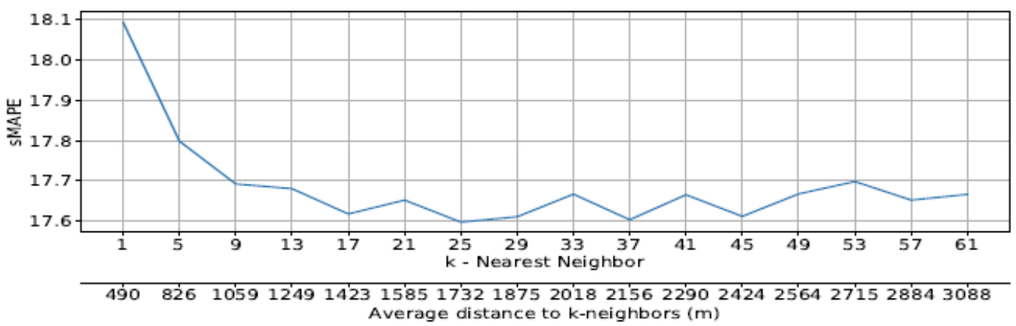
- Error distribution on two dataset
  - Models work better with “clean” dataset (D1)
  - Lower variation in results with the full-dataset (D2)

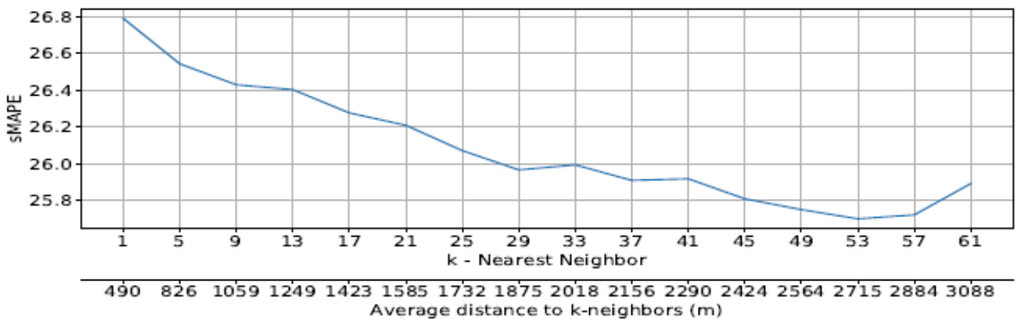


Comparing error distributions for different prediction methods for dataset D1 (left) and D2 (right)

Results

- Impact of neighbourhood size - k
  - In general, the size of neighborhood has effect to the prediction performance
  - Should set a threshold for k





Evaluating the impact of neighborhood on the predictive performance of neighborhood-augmented LSTM model on: D1 (left) and D2 (right)

Conclusion and outlook

- We propose a neighborhood-augmented LSTM model
  - Consider k-neighbors for prediction
  - Increase the performance of LSTM model
- Future work
  - Learn locally per stand and re-tune globally the predictions in the city
  - Including other sources of information (POI, events, traffic pattern...)

Thank you for your attention!
Questions?

 tai@l3s.de