Area-based mean speed estimation focusing on loop and sparse trajectory data

F. Ge[a], M. Zargayouna[b], A. Loder[c], L. Leclercq[a,*]

[a] Univ. Gustave Eiffel, ENTPE, LICIT-ECO7, Lyon, France
fei.ge@univ-eiffel.fr, ludovic.leclercq@univ-eiffel.fr
[b] Univ. Gustave Eiffel, IFSTTAR, COSYS-GRETTIA, Marne-la-Vallée, France
mahdi.zargayouna@univ-eiffel.fr
[c] Technical University of Munich (TUM), Munich, Germany
allister.loder@tum.de
* Corresponding author

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: spatial mean speed estimation; loop detector data; probe data; fusing traffic data

1 INTRODUCTION

As cities around the globe continue to grow and evolve, managing urban traffic becomes an increasingly complex challenge. In this context, traffic speed estimation is fundamental for effective traffic management. It is essential for alleviating congestion, enhancing road safety, optimizing traffic flow, and improving urban transportation systems' sustainability. Research on speed estimation is an indispensable and vital component for the operation of the intelligent transportation system.

Many research challenges remain in this field. One significant issue is the high data quality requirements of many data-driven approaches, such as the demand for high-resolution data and low missing rates. (Yang & Qian, 2019) list a comprehensive range of high-resolution datasets, including 5-minute interval speeds, traffic counts, incidents, weather, and local events, yet data availability and scarcity remain persistent challenges in speed estimation. Furthermore, much of the existing research is geographically limited to small networks or single highways. For instance, speed estimation in (Zhang, *et al.*, 2020) and (Kim, *et al.*, 2024) focus primarily on highway corridors. Even though a broader network was considered in (Yu & Gu, 2019), link-based methods require higher quality datasets and entail more significant computational costs and resources. Conversely, (Lopez, *et al.*, 2017) use the mean speed of the clustered region to represent single link speed for travel time estimation, proving the adequacy of area-based approaches. It can also help mitigate data sparsity issues to some extent. Additionally, the current use of spatiotemporal relationships in traffic data is underdeveloped. (Yang & Qian, 2019) employs data of all road segments across different times of the day as spatiotemporal features without a deeper exploration of these relationships. Therefore, this study aims to build a daily area-based speed estimation system for a large-scale network with extremely sparse data by Random Forest (RF) regression. A spatiotemporal data encoder using Graph Convolutional Networks (GCN) and Gated Recurrent Unit (GRU) models has been constructed to capture the spatiotemporal interdependencies within traffic data. Moreover, to address the challenge of extreme data scarcity, we try to construct a system that primarily utilizes abundant loop detector data, which can be robust and perform comparably to the system that relies on probe data only. Experiments are conducted to explore which is the best-aggregated description of the local observations from loops that can accurately capture the variations in the mean speed. OpenStreetMap (OSM), probe data and loop detector data are used in the experiments focusing on the city center of Munich, Germany.

2 METHODOLOGY

2.1 Step 1: data preprocessing

2.1.1 Network and traffic data preprocessing

The network is partitioned into simplified square grid units, referred to as a 'cell' in our study. The cells are defined as grid-area-based units measuring 1 km by 1 km, unless specified otherwise, corresponding to specific time slots. The effects of varying cell sizes will be discussed in Section 3.

The raw probe data consists of individual GPS trajectories that contain significant noise. For instance, the initial and final segments often include parking-related movements, sporadic noise points that deviate substantially from the normal trajectory, and brief stops for parking during the trip. Therefore, all the trajectories are cleaned dedicatedly to eliminate the aforementioned types of noise.

The presence of successive zeros is always a challenge when distinguishing between the absence of vehicles and null values in loop detector data. Consequently, sequences of zero values (more than three consecutive zeros) in the flow data have been removed from each detector to address this issue.

2.1.2 Input features extraction

The input features for speed estimation in this study are categorized into three main types: intrinsic features, loop detector features, and probe features.

Intrinsic features are acquired from the network information from OSM, which are attributes that are constant over time and solely related to the geographical location of the cell:

- ***Average speed limitation:*** The link-length weighted average of links' speed limits in the cell.
- ***Average road class:*** The average of the road class in the cell.
- ***Major roads:*** A binary indicator reflecting the presence of a major road (motorway) in the cell.
- ***Sum of betweenness centrality:*** The value to quantifies the centrality of nodes (intersections in the cell), characterizing the density and connectivity of the network within the cell.

For the loop detector features, this study employs commonly used descriptive statistics to analyze traffic flow and occupancy. These metrics include the ***average flow, standard deviation of flow, first quartile of flow, median of flow, third quartile of flow, average occupancy, and standard deviation of occupancy***. Each feature is calculated using data from detectors located within the specified cell.

Regarding the probe features, two primary attributes are analyzed to provide insights into the traffic dynamics surrounding the cell, enhancing the understanding of its traffic characteristics:

- ***Nearest speed:*** The speed of the closest trajectory to the cell that does not traverse the cell itself.
- ***Nearest distance:*** This measures the shortest distance from the cell to the nearest trajectory.

2.1.3 Spatial mean speed calculation

The mean speed for each cell at one specific time slot (15-minute interval) is defined as total travel distance divided by total travel time:

$$S_{j,t} = \frac{\sum_{i=1}^{n} td_{i,j,t}}{\sum_{i=1}^{n} tt_{i,j,t}} \tag{1}$$

Where $S_{j,t}$ represents the spatial mean speed of cell j at time t, $td_{i,t}$ denotes the travel distance of the i^{th} trajectory at time t within cell j, $tt_{i,t}$ is the travel time of the i^{th} trajectory at time t in cell j,n is the total number of trajectories at time t in cell j.

2.2 Step 2: spatiotemporal data encoding

GCN is a powerful neural network architecture designed to process data structured in graphs, utilizing convolutional techniques to effectively capture the graph's topological structure. On the other hand, GRU is a type of recurrent neural network optimized for sequence prediction tasks. In this setup, each cell map at a time interval is initially transformed into a graph structure, where each cell is treated as a node with connections to adjacent cells. The input features for each cell are transformed as the attributes for each node correspondingly. Subsequently, the input features are fed into both the GCN and GRU models to encode the spatiotemporal characteristics of the data.

2.3 Step 3: random forest regression

Random Forest is an ensemble machine learning method applicable to both regression and classification tasks. It builds upon the foundation of decision trees by aggregating multiple trees to form a "forest." The response variable $Y = y_1, y_2, y_3, \dots, y_n$ (spatial mean speed) is estimated by $X = x_1, x_2, x_3, \dots, x_n$ (each x is a set of explanatory variables as outlined in section 2.1.3). The training process of RF is structured into three primary steps: (1) Randomly select n samples from the training set data; (2) Train a regression tree on each of these samples; (3) Average the prediction results from all the trees.

3 SPEED ESTIMATION PERFORMANCE

This study conducts experiments focused on the city center of Munich, Germany, utilizing three distinct datasets. The first dataset, derived from OpenStreetMap (OSM), is used to construct the road network and to extract intrinsic environmental features. This dataset covers approximately 140 km^2 and includes about 16,000 links used for driving. The second and third datasets, comprising two months of probe and loop detector data, respectively, are provided by the Technical University of Munich (TUM), beginning in September 2022. The probe data includes 15,967 individual GPS trajectories, each tagged with a track ID, GPS coordinates, and timestamps. Meanwhile, the loop detector data includes readings from 5,386 detectors, which record traffic flow and occupancy information at 15-minute intervals. Thus, considering the data characteristics, the speed is calculated for each region unit every 15 minutes.

The study encompasses a total of 140 grid regions, each observed at 15-minute intervals from 7 AM to 9 PM, resulting in a total of 478,240 cells. However, only 10,836 of these cells contain valid speed and input features suitable for training, indicating an extreme missing data rate of 97.7%. This significant data scarcity substantially impacts the training process and contributes to larger errors in general. The performance of the trained model is evaluated using several metrics: Root Mean Squared Error (RMSE), Normalized Root Mean Squared Error (NRMSE), and the coefficient of determination (R^2).

Speed estimation has been performed on different datasets using different selections of input features. The results are shown in Table 1.

Probe data only: Use intrinsic features and probe features in speed estimation. Disregarding the data scarcity, probe data are generally the most reliable for estimating speeds. Thus, it's essential first to establish a baseline using only GPS data. This allows us to measure the additional value provided by loop detectors and to assess how well the loop detectors perform solely.

Fusion dataset: Use intrinsic features, probe features, and loop detector features to estimate speed. With the most information, fusion dataset should have the best performance in estimating speed. Compared with "probe data only", the increase caused by adding loop features can be evaluated.

Loop detector data only: Use intrinsic features and different combinations of loop detector features for the speed estimation. Considering the pros and cons, GPS data, while dependable, are often sparse and challenging to collect. Conversely, loop detector data are commonly available and abundant, yet they lack precision and are insufficient for accurate speed estimation. Therefore, it is worthwhile to investigate how to optimize speed estimation that can approximate GPS data quality, particularly in scenarios where loop detector data are abundant, but GPS data are not available. Seven loop features are proposed in this study; it is also important to find the best combination of the local observations from loops that can accurately capture the variations in the mean speed. Three combinations of the loop features are tested: (1) Combination 1: ***average flow*** and ***standard deviation flow***; (2) Combination 2: ***average flow, standard deviation of flow, first quartile of flow, median of flow, third quartile of flow,*** (3) Combination 3: all the seven features.

Table 1 – *Speed estimation performance using different input features*

Cell size	Dataset	*RMSE (m/s)*	*NRMSE*	*R^2*
0.8 km * 0.8 km	Probe data only	3.608	0.330	0.574
	Fusion dataset	3.539	0.333	0.590
	Loop only (combination 1)	3.928	0.369	0.495
	Loop only (combination 2)	3.865	0.363	0.511
	Loop only (combination 3)	3.870	0.364	0.510
1 km * 1 km	Probe data only	3.571	0.342	0.512
	Fusion dataset	3.433	0.329	0.551
	Loop only (combination 1)	3.838	0.367	0.438
	Loop only (combination 2)	3.790	0.363	0.452
	Loop only (combination 3)	3.802	0.364	0.449
1.2 km * 1.2 km	Probe data only	4.006	0.355	0.597
	Fusion dataset	3.815	0.338	0.635
	Loop only (combination 1)	4.507	0.399	0.491
	Loop only (combination 2)	4.423	0.392	0.509
	Loop only (combination 3)	4.388	0.389	0.517

4 DISCUSSION

This study utilizes a spatiotemporal data encoder and Random Forest regression for area-based mean speed estimation across a large-scale network, utilizing extremely sparse data. Experimental analyses in Munich across various datasets and input features reveal that 1-km-sized cells incorporating all flow-related features yield the best possible accurate speed estimations in the absence of probe data among all the scenarios proposed in this study. The inclusion of occupancy-related features does not consistently enhance performance, even though additional information is provided. However, due to time limitations and technical issues, the temporal encoding requires further refinement. Detailed results and additional refinements will be presented at the conference.

References

Kim, Y., Tak, H.-y., Kim, S. & Yeo, H. (2024). A hybrid approach of traffic simulation and machine learning techniques for enhancing real-time traffic prediction. *Transportation Research Part C: Emerging Technologies*, 160, p. 104490.

Lopez, C., Leclercq, L., Krishnakumari, P., Chiabaut, N. & Van Lint, H. (2017). Revealing the day-to-day regularity of urban congestion patterns with 3D speed maps. *Scientific Reports*, 7(1), p. 14029.

Yang, S. & Qian, S. (2019). Understanding and Predicting Travel Time with Spatio-Temporal Features of Network Traffic Flow, Weather and Incidents. *IEEE Intelligent Transportation Systems Magazine*, 11(3), pp. 12-28.

Yu, J. J. Q. & Gu, J. (2019). Real-Time Traffic Speed Estimation With Graph Convolutional Generative Autoencoder. *IEEE Transactions on Intelligent Transportation Systems*, 20(10), pp. 3940-3951.

Zhang, Z., Yuan, Y. & Yang, X. (2020). A Hybrid Machine Learning Approach for Freeway Traffic Speed Estimation. *Transportation Research Record,* 2674(10), pp. 68-78.