Data Fusion: Traffic Flow Estimation Based on a GNN Approach for Loop Detector Data

Leveraging Spatio-Temporal Features and Floating Car Data – A work in progress

Amir Babaei,
Prof. Dr. Tobias Hagen

European Transport Conference 2025

Frankfurt University of Applied Sciences
amir.babaei@fb3.fra-uas.de

Why Accurate Traffic Data is Crucial for ITS

Accurate traffic state variables—flow, speed, and density—are the bedrock of modern Intelligent Transportation Systems (ITS) [1]. They directly enable:

Real-Time Operations

- Adaptive traffic signal control
- Incident detection & management [2]
- Dynamic route guidance [3]

Strategic Planning & Analysis

- Identifying congestion bottlenecks [4]
- Infrastructure investment planning [5]
- Environmental impact modeling [6]

Without a complete and accurate picture of the traffic state, these applications fail or become unreliable.

The Data Reality: Two Imperfect Sources

To understand the traffic state, we primarily rely on two data types, each with significant limitations:

Eulerian Data(Fixed Sensors)

- E.g., Loop Detectors [7]
- Provide point-based "ground truth".
- Problem: Suffer from frequent failures, creating spatial and temporal data gaps.

Lagrangian Data (Mobile Sensors)

- E.g., Floating Car Data (FCD) [8]
- Provides wide network coverage.
- Problem: A biased sample with unknown and variable penetration rates.

The core challenge is to fuse these imperfect sources to create a complete and accurate picture.

Related Work: Forecasting Methods & Their Limits

Classical approaches struggle with the complex nature of traffic networks:

- **Simulation-based Models** (e.g., micro/meso/macro) [9] *Limit:* Often computationally expensive for large networks or they oversimplify vehicle behavior.
- **Deterministic Models** (e.g., Kinematic Wave Theory) [10] *Limit:* Struggle to capture the stochastic, non-linear dynamics of real-world traffic.
- **Statistical & Classic ML** (e.g., ARIMA, SVR) [11, 12] *Limit:* Often treat sensors as independent time-series, failing to model the crucial spatial relationships between them.

The fundamental limitation

Traffic is inherently a **graph**—a network of interconnected sensors. Models that ignore this topology will have limited success.

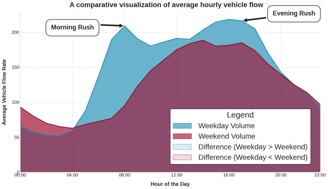
Presentation Outline

This presentation of our work-in-progress is structured in three parts:

1. **Data Pipelines & Ground Truth Creation**
   Detailing the study area, data sources (LDD, FCD), and the fusion process.
2. **Analysis of FCD Penetration Rate Dynamics**
   Investigating the key factors that influence the representativeness of FCD.
3. **GNN-based Data Imputation**
   Presenting the spatio-temporal model design, experimental setup, and key results for filling data gaps.

Data Pipeline I: Loop Detector Data (LDD)

Map of the study area and loop detectors



Average hourly traffic flow; typical weekday vs. a weekend day.

Source: High-Resolution LDD [13]

- More than 500 detectors, providing minute-by-minute flow (q) and occupancy (o).
- Serves as our ground truth for total vehicle counts (N_{Total}).

Pre-processing Pipeline:

- Flagging of invalid data based on physical constraints (e.g., $o \notin [0, 100]$, $q > C_{max}$).
- Removal of detectors with less than 5% data availability.
- Imputation and aggregation to 10-minute intervals.

Data Pipeline II: Floating Car Data (FCD)

Source: Commercial FCD [14]

- ≈ 19M raw trips
- ≈ 3M unique devices
- Provides probe counts (N_{FCD})

Preprocessing Pipeline:

- Geospatial filtering
- Large-scale map-matching to OSM using Valhalla [15, 16]

Raw FCD trajectories (left) vs. map-matched network volume (right).

Data Pipeline III: Fusion for Ground Truth Analysis

To analyze FCD representativeness, we fuse the two data streams at co-located points on the OSM network.

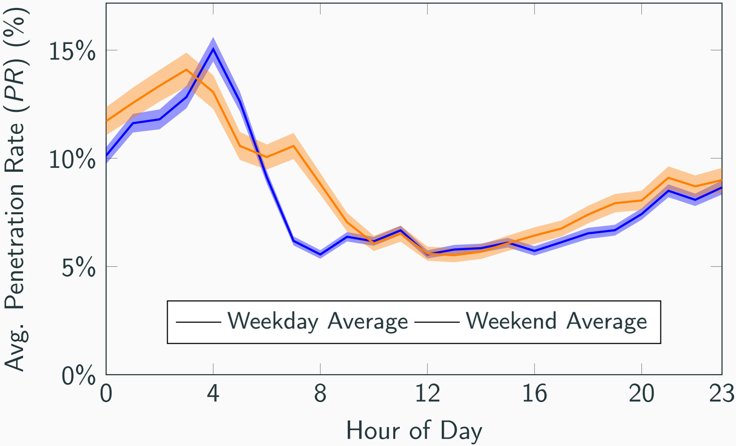
- LDD sensors are spatially mapped to their corresponding OSM road segments.
- FCD trips are counted as they traverse these same segments.

This allows us to calculate the empirical **Penetration Rate (PR)** for any location l at time t:

$$PR(l,t) = \frac{N_{FCD}(l,t)}{N_{Total}(l,t)}$$

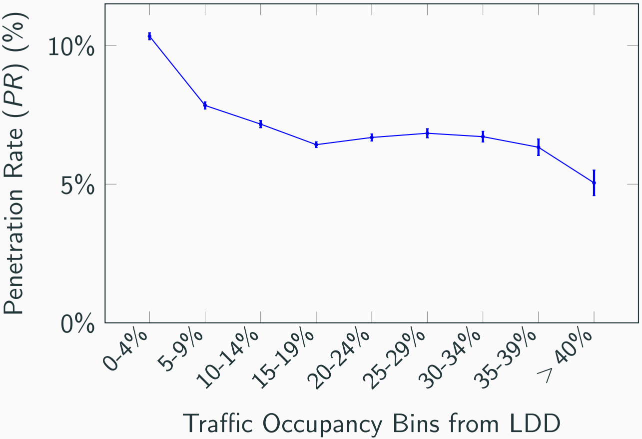
This fused dataset is the foundation for both understanding FCD bias and for training our GNN.

FCD Analysis I: Temporal Dynamics of Penetration Rate



- While both day types follow a similar daily pattern, with peaks in the early morning and evening, **weekends** show a significantly **higher PR** during the day.

FCD Analysis II: Impact of Congestion on PR



- Increased traffic congestion causes a significant drop in the Penetration Rate.
- This is likely due to the traffic avoidance behavior of taxi and fleet drivers.

GNN Model Design for Data Imputation

To capture the network's spatio-temporal nature, we designed a GATGRU model.

1. **Graph Representation**
   - Sensors are nodes (V).
   - Connectivity forms edges (E).
   - Edge weights are a Gaussian kernel on distance:

$$A_{ij} = \exp\left(-\frac{\text{dist}(i,j)^2}{\sigma^2}\right)$$

2. **Input Features**
   - LDD history (flow, occ, speed)
   - Temporal Embeddings (time of day, day of week)

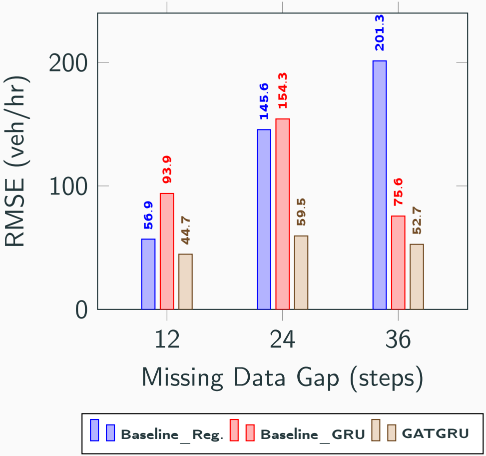
3. **GATGRU Architecture**
   - A **Graph Attention (GAT)** layer aggregates spatial information from important neighbors.
   - A **Gated Recurrent Unit (GRU)** layer processes the output as a time-series to learn temporal patterns.

GNN Experiment: Setup & Flow Imputation Results

Setup

- **Task:** Impute artificially created data gaps in the test set.
- **Models Compared:**
  - ‘Baseline_Reg.‘: Statistical
  - ‘Baseline_GRU‘: Temporal
  - ‘GATGRU‘: Spatio-Temporal
- **Scenarios:** 2, 4, and 6-hour gaps (12, 24, 36 steps).
- **Metric:** RMSE

Flow Imputation (RMSE)

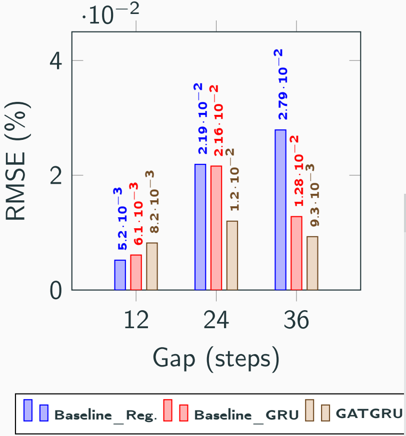


Insight: The spatio-temporal GATGRU model consistently and significantly outperforms both baselines, especially for longer gaps.

GNN Experiment: Occupancy Results & Performance Gain

The GATGRU model's advantage becomes more pronounced as the imputation task becomes more difficult.

Occupancy Imputation (RMSE)



GATGRU Advantage over Best Baseline (Flow RMSE)

- At 2-hour gap: **21%** lower error
- At 4-hour gap: 59% lower error
- At 6-hour gap: 74% lower error

Key Takeaway

The GNN's ability to leverage spatial information from neighboring sensors provides a massive advantage when local temporal data is missing for extended periods.

Discussion

Why does the GNN-based fusion approach work?

- **Spatial Learning (GAT):** The model learns which neighboring sensors are most influential.
- **Temporal Learning (GRU):** The model captures recurring patterns (e.g., rush hours), allowing it to make context-aware predictions.
- **Feature Fusion Value:** The model is enhanced by exogenous features. Temporal embeddings provide crucial time context.

- **FCD such as counts and speed can provide an "independent witness" signal, and could be potentially beneficial for model accuracy**

Applications & Future Work

Our trained models enables several practical applications:

- Enables investigating the features changing PR
- Estimation of traffic flow on street level granularity
- Enhanced inputs for emission models and traffic models.

Future Work:

- Directly integrate FCD counts and speed as a feature to quantify its impact.
- Explore features proven to be important in PR estimation

Conclusion

1. **Problem:** Missing data from physical sensors is a significant barrier for effective Intelligent Transportation Systems.
2. Our experiments demonstrate that FCD and LDD have variable spatial and temporal relations.
3. We proposed a spatio-temporal GNN model that learns from the sensor network structure and temporal patterns to impute missing data.

References

[1] **Lily Elefteriadou**. "Flow, Speed, Density, and Their Relationships". In: (2014). DOI: 10.1007/978-1-4614-8435-6_3.

[2] **B. S. Kerner et al**. "Traffic state detection with floating car data in road networks". In: *2005 IEEE Intelligent Transportation Systems, 2005. Proceedings*. 2005, pp. 700–705. DOI: 10.1109/ITSC.2005.1520133.

[3] **A. Simroth and H. Zähle**. "Travel Time Prediction Using Floating Car Data Applied to Logistics Planning". In: *IEEE Transactions on Intelligent Transportation Systems* 12.1 (2011), pp. 243–253. DOI: 10.1109/TITS.2010.2090521.

[4] **O. Altintasi, H. Tuydes-Yaman, and K. Tuncay**. "Detection of urban traffic patterns from Floating Car Data (FCD)". In: *Transportation Research Procedia* 22 (2017), pp. 382–391. DOI: 10.1016/j.trpro.2017.03.057.

[5] **F. Su et al**. "Long-term forecasting oriented to urban expressway traffic situation". In: *Advances in Mechanical Engineering* (2016). DOI: 10.1177/1687814016628397.

[6] **Y. Jiang et al**. "Estimation of Hourly Traffic Flows from Floating Car Data for Vehicle Emission Estimation". In: *Journal of Advanced Transportation* 2021 (2021). DOI: 10.1155/2021/6628335.

[7] **W. Weijermars and E. C. van Berkum**. "Detection of Invalid Loop Detector Data in Urban Areas". In: *Transportation Research Record*. 2006. DOI: 10.3141/1945-10.

[8] **T. Wang and J. Hu**. "Applying floating car data in traffic monitoring". In: *International Conference Control Science and Systems Engineering*. 2014. DOI: 10.1109/CCSSE.2014.7224516.

[9] **S. Druitt**. "An introduction to microsimulation". In: *Traffic Engineering and Control* (1988).

[10] **D. C. Gazis, R. Herman, and R. Rothery**. "Analytical methods in transportation: mathematical car-following theory of traffic flow". In: *Journal of Engineering Mechanics-Asce* (1963). DOI: 10.1061/JMCEA3.0000437.

[11] **J. Wang et al**. "Research on short-term traffic flow prediction based on SARIMA model". In: *Sixth International Conference on Electromechanical Control Technology and Transportation (ICECTT 2021)*. 2022. DOI: 10.1117/12.2623999.

[12] **J. Petrlik and L. Sekanina**. "Towards Robust and Accurate Traffic Prediction Using Parallel Multiobjective Genetic Algorithms and Support Vector Regression". In: *International Conference on Intelligent Transportation Systems*. 2015. DOI: 10.1109/ITSC.2015.360.

[13] **Mobilithek**. *Die Nationale Mobilithek*. https://mobilithek.info/. Accessed: September 15, 2025. 2025.

[14] **INRIX**. *INRIX - The Global Leader in Transportation Data and Analytics*. https://inrix.com/. Accessed: September 15, 2025. 2025.

[15] **Valhalla**. *Valhalla - Open Source Routing Engine*. https://github.com/valhalla/valhalla/. Accessed: September 15, 2025. 2025.

[16] **Siavash Saki and Tobias Hagen**. "A Practical Guide to an Open-Source Map-Matching Approach for Big GPS

Thank You

Questions?