Promotionszentrum
Mobilität und Logistik

FRANKFURT
UNIVERSITY
OF APPLIED SCIENCES

Exposé for the Doctoral Project

Working Title: Graph Neural Networks for Estimating Traffic State Variables

submitted by

M.Sc. Amir Babaei

Frankfurt University of Applied Sciences
Research Lab for Urban Transport (ReLUT)

Frankfurt am Main, Germany
amir.babaei@ieee.org

Submitted on: July 3, 2026

1st Supervisor:	Prof. Dr. Tobias Hagen, Frankfurt University of Applied Sciences
2nd Supervisor:	Prof. Dr.-Ing. Bernhard Friedrich, Technische Universität Braunschweig

Contents

1 Research Context and Motivation

Effective (road) traffic management and the assessment of transport policies require accurate, current as well as historical data on the "state of the system" across the entire urban (road) network. Traditionally, this relies on Stationary Detector Data (SDD), such as inductive loops and traffic cameras. While SDD provides accurate "ground truth" on counts and occupancy at specific locations, it is infrastructure-heavy, expensive to maintain, and spatially sparse, usually leaving most of the road network unobserved (Ashwini and Sumathi 2020).

From about the mid-2000s, companies and fleet operators began delivering “probe” or floating-car data (FCD - consisting of GNSS[1] data) commercially (HERE, TomTom, INRIX, etc.). FCD converts vehicles into ubiquitous, roaming sensors across the road network (Turksma 2000). Over the following decade(s), these datasets matured into near-real-time, wide-coverage sources that let policy-makers and researchers measure variables at network scale without installing hardware on every road (Zhang et al. 2024).

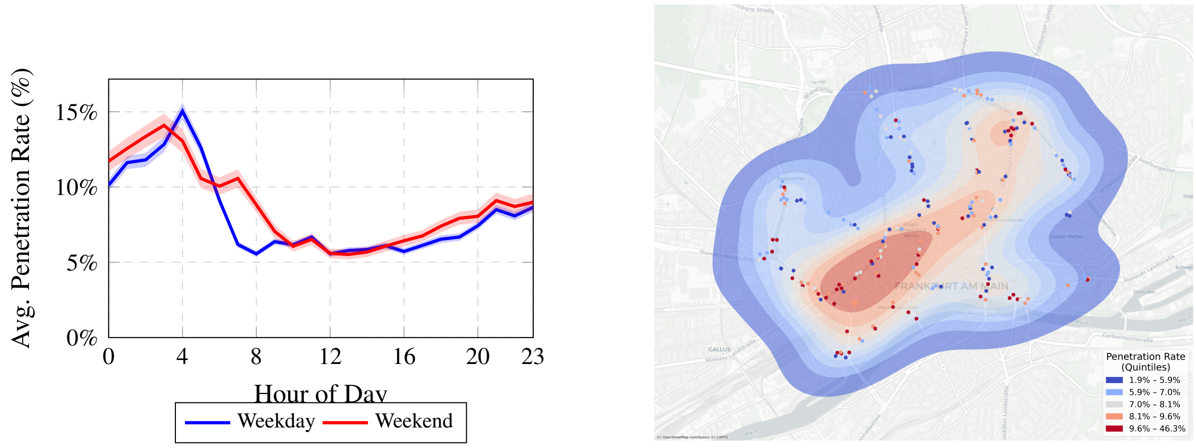
Within the FCD, the trips can be categorized into the following four categories with respect to a certain area (e.g. the boundaries of a city): (1) internal flows (within the city), (2) origin flows (starting in the city and ending outside the city), (3) destination flows (starting outside the city and ending in the city), and (4) transit flows (starting and ending outside but passing through the city). This allows analyzing the origins and destinations as well as the trajectories in between. Hence, route choice and driving style (speed, acceleration) can be analyzed. This also makes emission calculation more reliable.

While FCD is already frequently used in practice to measure the average speed of traffic, a critical methodological barrier prevents its use for vehicle counting: The "Penetration Rate" (PR)—the ratio of the number of vehicles included in the FCD to total number of actual vehicles—is unknown and highly dynamic. The PR varies significantly across space and time depending on congestion levels and urban topology (Zhang et al. 2024) and, of course, on the respective FCD data provider and the data sources contained therein.

Consequently, treating the PR as a static scalar across the network and using this as an upscaling factor the FCD introduces a significant bias. Preliminary large-scale empirical analysis was conducted by the author in Frankfurt am Main based on the FCD by the private company INRIX which was analyzed against a network of around 600 loop detectors providing minute-by-minute flow (q) and occupancy (occ). The study quantifies PR variability, revealing a Coefficient of Variation (CV) of 0.75 and location-specific rates ranging drastically from 1.9% to 46.2%—far from the network-wide mean of 8.31%. The analysis highlights two critical dynamics: First, the PR is inversely correlated with congestion, dropping to $\approx 5.6\%$ during morning peak hours on weekdays while rising above 15% in free-flow conditions (see Figure 1a). Second, it exhibits distinct spatial heterogeneity, with high concentrations near transport hubs (e.g., Central Station) and suppressed rates in business districts, likely due to the specific composition of the fleet (see Figure 1b). Accurate upscaling of FCD to total flows requires data-driven methods that are able to learn these complex, nonlinearities.

This doctoral project aims to bridge this gap by researching whether a **Physics-Informed Graph Neural Network** developed for this purpose, based on the fusion of different data sources, is able to

[1] Global Navigation Satellite System (GNSS) is the standard generic term for satellite navigation systems that provide autonomous geo-spatial positioning with global coverage. The Global Positioning System (GPS) is one specific, widely known type of GNSS.



(a) Temporal Dynamics of FCD Penetration Rate

(b) Spatial Heterogeneity of PR

Figure 1: Empirical Analysis of FCD Penetration Rates in Frankfurt: (a) Diurnal PR patterns for weekdays vs. weekends (shaded areas denote 95% CI); (b) Spatial distribution heatmap overlaid with sensor locations by PR quintile. Map data ©OpenStreetMap contributors.

create accurate estimates. By fusing the precise but sparse ground truth of SDD with the broad but potentially incomplete FCD, the model seeks to learn the complex dynamics of the penetration rate. This will effectively turn every street segment covered by FCD into a "virtual counting station," enabling network-wide traffic state estimation without the need for dense hardware infrastructure.

As explained in detail in the next section, graph neural networks (GNNs) have been predominantly utilized for short-term (so-called "spatio-temporal") forecasting for streets / areas across the street network (graph) for which historical SDD is available. FCD is in this context a source for useful variables/features which may raise the forecasting quality of the model. However, it is much less common to use GNN to estimate traffic conditions for streets for which there is no historical SDD available, which is in the literature referred to as "network-wide estimation" or "spatial inference". By leveraging the GNN's ability to learn structural spatial and temporal dependencies, it is possible to generalize patterns from data-rich areas to estimate the state of the entire road network. Integrating FCD trajectories in network-wide estimation may be particularly fruitful, as FCD exist for areas where SDD is not available. Even more information that helps for forecasting and network-wide estimation can be derived by integrating physical constraints of streets / street segments summarized by the Macroscopic Fundamental Diagram (MFD). In Section 2 the implied research gaps will be explained in greater detail.

The relevance of this research goes far beyond enabling public authorities to observe vehicle flows in every street segment:

- Cities frequently implement transport policies (e.g., low-emission zones, traffic calming, on-street parking regulations) to improve public health and safety. Currently, planners rely on simulations for "ex-ante" planning but usually lack the granular data for "ex-post" evaluation of what actually happened (Zhang et al. 2024). This research provides the high-resolution flow and density data necessary to accurately assess the real-world impact of these policies on emissions and congestion.

- The installation and maintenance of inductive loop detectors or camera systems incur high costs for municipalities(Ashwini and Sumathi 2020). By developing a method to upscale existing FCD, this research offers a scalable, low-cost alternative that reduces the dependency on dense physical infrastructure while maximizing the value of existing commercial data assets. FCD will remain an important research data source in the future since autonomous vehicles generate GNSS-based trajectories as they are needed for navigation, safety, AI training, and system monitoring.

- The insights to be gained are not limited to street networks and cars, but can also be transferred to other means of transport and other networks, for example, bikes and public transport.

2 Previous Research I: Evolution of Traffic State Estimation Methods

The methodology for estimating and imputing traffic states has undergone a significant transformation, evolving from linear statistical models to high-dimensional tensor operations, and finally to physics-aware, non-Euclidean deep learning architectures. Before describing these methodological developments, I will first clarify the dimensions of the estimation problem in the next section.

2.1 Dimensions of Estimation: Imputation, Forecasting, and Spatial Inference

Traffic State Estimation (TSE) infers variables like flow, density, and speed from partially observed data (Seo et al. 2017). It can be categorized along a space-time matrix, as visualized in Figure 2.

1. **Imputation (Restoration):** The recovery of missing values at locations equipped with sensors (e.g., due to device failure or transmission errors). This task relies on historical patterns and spatial correlations from immediate neighbors to fill "holes" in the observation matrix.

2. **Forecasting (Prediction):** The prediction of future states ($t+h$) for locations with historical data. This is the traditional domain of time-series analysis and Recurrent Neural Networks, utilizing past lags to project short-term trends.

3. **Network-wide estimation / Spatial Inference** involves the estimation of states for road segments that have *never* been equipped with sensors (the "Virtual Station"). This is the most complex challenge, as there is no historical ground truth for the model to train on directly. The model must learn to generalize the underlying physics or relationships from observed locations (graph) and transfer this logic to unobserved parts of the network.

As explained in the following, most existing GNN approaches focus on Tasks 1 and 2, utilizing the graph structure to improve forecasting accuracy. However, Task 3—network-wide estimation / spatial inference—remains under-explored, primarily because standard data-driven models struggle to generalize to edges where the target variable (volume) is completely absent from the training dataset.

2.2 The Statistical Era: Linear and Temporal Focus

Early research relied heavily on parametric statistical models to interpret traffic variables as univariate time series. The **Auto-Regressive Integrated Moving Average (ARIMA)** model, firstly applied to traffic by Ahmed and Cook (1979), established a foundation for capturing temporal trends and seasonality. While effective for short-term forecasting on single road segments under stable conditions, these linear models are not able to capture the complex, non-linear spatiotemporal dependencies and rapid state transitions inherent in large-scale urban networks (Afandizadeh et al. 2024). Concurrently, the **Kalman**

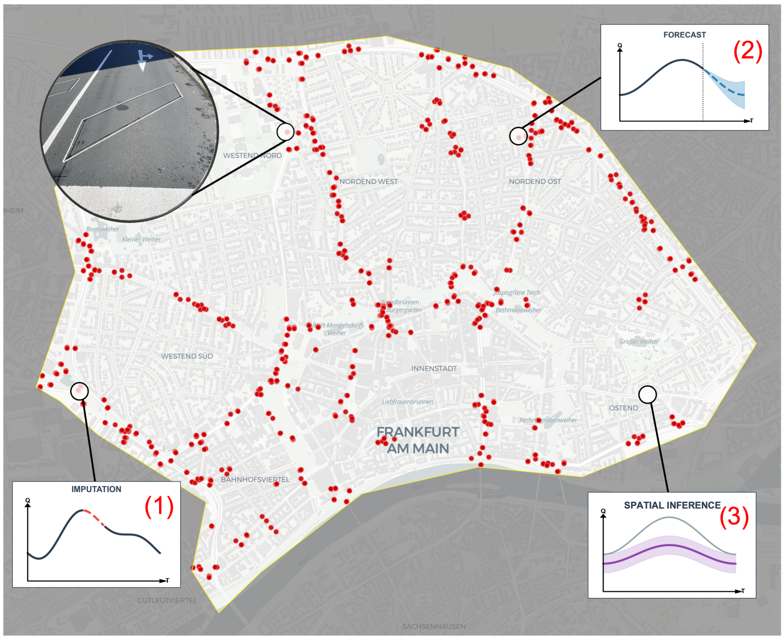


Figure 2: An example of stationary detectors spread in a city that measure traffic state; Frankfurt am Main Alleenring. *Imputation* fills gaps in historical data (Task 1). *Forecasting* predicts future states for observed locations (Task 2). *network-wide estimation / spatial inference* (the focus of this work) infers states for locations with no physical infrastructure (Task 3).

Filter became the standard for real-time state estimation due to its recursive ability to update states with new noisy measurements. However, its optimal performance is strictly bound to linear system assumptions, limiting its effectiveness in highly dynamic environments (Afandizadeh et al. 2024). These models share the common feature that the time series for each street (segment) are treated separately. That means spatial interdependencies are not taken into account. Moreover, the early models were based on univariate time series without considering explanatory variables (such as the weather conditions, etc.). This implies that they can only be used for Task 1 (Imputation) and Task 2 (Forecast), but not for Task 3 (network-wide estimation / spatial inference).

2.3 The Matrix and Tensor Era: Capturing Global Correlations

To address the challenge of missing data in network-wide estimation, researchers shifted towards modeling traffic data as high-dimensional structures, taking spatial interdependencies into account. **Matrix Factorization (MF)** and **Tensor Decomposition (TD)** techniques such as Bayesian Tensor Decomposition, introduced by Chen et al. (2019), emerged as powerful solutions for imputation, representing traffic data as multi-dimensional tensors (e.g., Location $\times$ Time $\times$ Day). These methods impute missing values by capturing multi-dimensional correlations, such as day-to-day recurrence patterns while accounting for all locations simultaneously (Zhang et al. 2024). While tensor structures outperform matrix representations in preserving information within large-scale historical datasets, they are typically transductive; meaning they lack the ability to generalize to new data streams without computationally expensive retraining (Zhang et al. 2024).

2.4 The Deep Learning Era: Non-Linearity and Graph Topology

The advent of deep learning introduced models capable of capturing complex non-linear relationships that statistical and tensor methods miss.

- **Recurrent Neural Networks (RNNs):** Architectures such as **Long Short-Term Memory (LSTM)** and **Gated Recurrent Units (GRU)** overcame the limitations of traditional RNNs, enabling the effective learning of long-term temporal dependencies (Afandizadeh et al. 2024). These models have proven highly effective for sequential traffic prediction but inherently treat spatial data as independent sequences, often neglecting the crucial topological interactions between connected road segments (Zhang et al. 2024).

- **Graph Neural Networks (GNNs):** To explicitly model the non-Euclidean topology of road networks, Graph Neural Networks (Kipf and Welling 2016) have become the current state-of-the-art(Jiang and Luo 2022). Pioneering frameworks include the **Spatio-Temporal Graph Convolutional Network (ST-GCN)** by Yu et al. (2018), which integrates graph convolutions with gated temporal convolutions, and the **Diffusion Convolutional Recurrent Neural Network (DCRNN)** by Li et al. (2017), which captures spatial dependencies using bidirectional random walks on graph nodes enabling the identification of upstream (e.g., congestion propagation) and downstream traffic patterns. Further advancements like **Graph WaveNet** by Wu, Pan, et al. (2019) introduced adaptive dependency matrices to capture hidden spatial structures without requiring a pre-defined graph. Crucially, **Graph Attention Networks (GATs)**, introduced by Veličković et al. (2017), allow models to learn dynamic edge weights, enabling the aggregation of information from neighboring nodes with varying importance based on real-time traffic states (Zhang et al. 2024). Furthermore, attention mechanisms are particularly effective at capturing dynamic daily and weekly periodicities in spatial-temporal traffic forecasting (Guo et al. 2019).

- **Physics-Informed Graph Neural Networks**: Recently, a new paradigm has emerged seeking to combine the data-driven power of GNNs with the theoretical rigor of traffic flow models Wang et al. (2024). This approach embeds physical laws directly into the learning process to prevent physically impossible predictions (e.g., negative flows). For instance, this enforces flow conservation constraints (Kirchhoff's Law), ensuring that the total volume of traffic entering a node equals the volume exiting it.

  Foundational to this approach is the **Fundamental Diagram (FD)**, established by Greenshields (1935), which defines the equilibrium between flow (q), density (k), and speed (v) via the identity $q = k \cdot v$. This macroscopic relationship was empirically validated at the network level by demonstrating the existence of urban-scale Macroscopic Fundamental Diagrams (MFDs) (Geroliminis and Daganzo 2008; Treiber and Kesting 2013). Furthermore, the temporal evolution of these states is governed by the conservation of vehicles, mathematically formalized in the **Lighthill-Whitham-Richards (LWR)** kinematic wave models (Lighthill and Whitham 1955; Richards 1956) and their modern discrete successor, the **Cell Transmission Model (CTM)** (Daganzo 1994).An empirical example of these macroscopic relationships is illustrated in Figure 3 constructed based observed data in Frankfurt am Main.

  Recent studies, such as Wang et al. (2024) and Pan et al. (2025), integrate these principles into

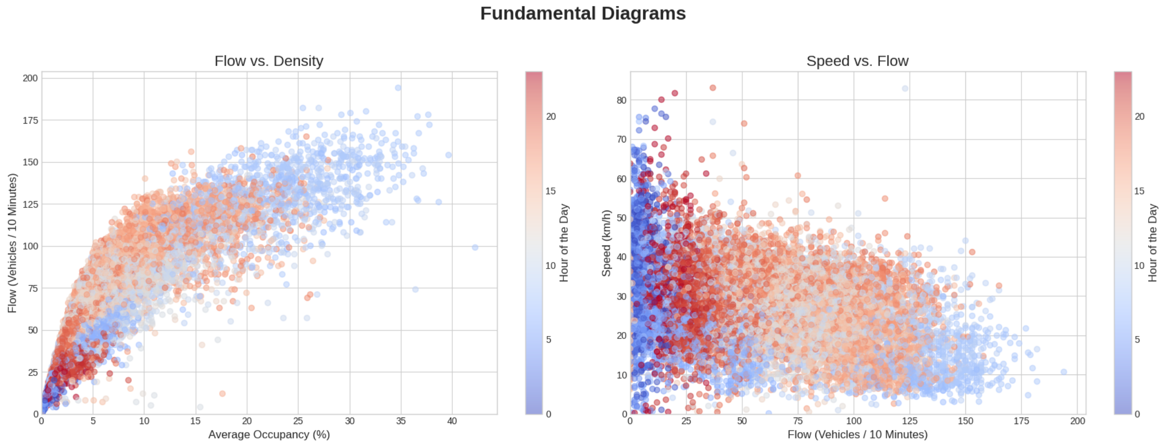


Figure 3: Visualization of real-world traffic states using multi-source data fusion on Hansaallee Street. The plots display Flow-Occupancy (left) and Speed-Flow (right) relationships aggregated over four months (Nov 2024 – Feb 2025) in Frankfurt am Main. space-mean-speed is derived from FCD trajectories, while flow and occupancy represent Loop Detector Data (LDD). Colors indicate the time of day.

deep learning architectures. Wang et al. (2024), for instance, developed a "Fundamental Diagram Learner" to constrain highway predictions, while Pan et al. (2025) utilized a stepwise framework where the GNN learns the residual errors of a physics-based model. However, approaches vary significantly, ranging from soft regularization (loss functions) to hard architectural constraints.

3 Previous Research II: Multi-Source Data Fusion

To overcome the spatial limitations of fixed sensors, recent literature has shifted towards Multi-Source Data Fusion, specifically integrating Stationary Detector Data (SDD) with Floating Car Data (FCD). Early network-scale estimations relied heavily on partial information from GPS-enabled taxi OD trips (Zhan et al. 2013), but modern methodologies aim to combine the complementary strengths of these heterogeneous sources: the high temporal resolution and precision of SDD with the extensive spatial coverage of FCD. Afandizadeh et al. (2024) further highlight that deep learning architectures benefit significantly from this multi-modal input, as it allows models to capture complex spatiotemporal dependencies that single-source approaches miss.

However, as highlighted in Section 1, the effectiveness of this integration is strictly bound by the quality of the probe data. Rizvi and Friedrich (2024) demonstrate that the derivation of accurate network-wide states (such as the MFD) is heavily contingent upon the Penetration Rate (PR).

4 Research Gaps and Questions

While the literature has established the utility of GNNs and the concept of Physics-Informed Learning, critical gaps remain regarding their optimal application to network-wide flow estimation and data fusion. This proposal addresses three specific limitations:

1. **Research Gap 1: The "Virtual Station" Spatial Inference Problem**
   The ultimate challenge in urban monitoring is task 3: estimating flows on road segments with

zero physical sensors. While task 1 (imputation on graphs, i.e., restoring missing values at sensor locations) is a known task addressed by approaches like Graph Markov Networks (Cui et al. 2020), network-wide estimation (creating "Virtual Counting Stations") requires a model that can generalize the learned relationship between FCD and Total Flow to links without any historical observation. This remains an open challenge due to the difficulty of verifying estimates where ground truth is absent.

→ **Research Question 1:** Can a graph-based framework fusing SDD and FCD accurately estimate total traffic flow at "Virtual Stations" (unobserved links) by generalizing the ground truth from observed locations?

2. **Research Gap 2: Underutilization of Contextual Information in Fusion**
   Several studies have implemented data fusion to combine FCD and SDD. However, these methods often used FCD just to calculate vehicle counts, neglecting the rich contextual information (e.g., trajectory patterns, temporal profiles) inherent in the data. (Zhang et al. 2024) Consequently, many relevant aspects known from traffic flow theory and previous studies to characterize the traffic situation are not taken into account.(Wang et al. 2024)

   → **Research Question 2:** How can contextual information extracted from FCD trajectories be leveraged in a data-fusion process to improve the accuracy of flow estimation? Which information improves the accuracy?

3. **Research Gap 3: Ambiguity in Physics-Informed Integration Strategies**
   Finally, to constrain these estimates in data-sparse regions, physical laws can be integrated. Although Physics-Informed GNNs (PI-GNNs) have recently been introduced (Wang et al. 2024; Pan et al. 2025), there is no established consensus on the methodology for integrating traffic flow theory into deep learning models for the specific task of network-wide estimation. Current approaches vary widely—from modifying loss functions to altering network architectures—and it remains unclear which strategy works best with regard to accuracy of estimates.

   → **Research Question 3:** How can physical relationships in traffic flow theory (e.g., Fundamental Diagram, conservation laws) be best leveraged within a GNN framework to maximize accuracy and physical consistency in traffic state estimation?

5 Methodology and Approach

This research proposes to gain insights into a generalizable framework for network-wide traffic state estimation. The methodology shall address the limitations of sparse sensor coverage and potential incompleteness of FCD by treating the road network as a graph and systematically investigating the optimal integration of traffic flow theory into deep learning models. The approach is structured into three core components: a data fusion pipeline which takes FCD contextual information into account, a comparative implementation of various physics-informed architectures, and a latent variable module for network-wide estimation.

5.1 Scalable Data Fusion and Graph Construction

The foundation of the framework is an automated Extract-Transform-Load (ETL) pipeline designed to fuse heterogeneous traffic data onto a unified graph structure $G = (V, E)$, with V representing the Vertices or "road segments" and E the Edges that connect them. This process is transferable to any urban network.

- **Map-Matching and Topological Integration:** Raw FCD trajectories (timestamped GNSS sequences) are map-matched to the underlying road network (e.g., OpenStreetMap) using existing algorithms (Saki and Hagen 2022). This step converts unstructured trajectory points into edge-specific attributes, assigning a dynamic sample of vehicle speeds and counts to each graph node (road segment).
- **Contextual Feature Extraction (Addressing RQ2):** This involves fusing the graph with auxiliary data sources to generate a feature vector for each node-time pair. Key features include:
  - *Vehicle Composition Context:* Features derived from provider metadata, such as Provider Type (Consumer vs. Fleet).
  - *Traffic Flow Context:* Metrics derived from trajectory analysis, including space mean speeds, turn-ratio at intersections and stop-and-go patterns (distinguishing moving trips from non-moving events).
  - *Semantic and Urban Context:* High-level features extracted from trajectory geometries and destination areas, such as the Urbanity Index (ratio of Euclidean to actual trip distance), Maneuver Point Density, and unsupervised City Segmentation Labels (e.g. industrial vs. office-related areas), to characterize the spatial environment and its influence on the dynamic penetration rate. (Hamann and Hagen 2025)
- **Spatio-Temporal Alignment:** Stationary Detector Data (SDD), serving as the ground truth for flow (q) and occupancy (occ), is spatially joined to the corresponding graph edges. The pipeline aggregates both data streams into consistent temporal bins, creating a feature matrix X_t containing both sparse ground truth and broad, context-rich probe measurements.
- **Graph Representation:** The road network is modeled as a directed graph where nodes V represent road segments and edges E represent physical connectivity. This topology is encoded into an adjacency matrix A, which could be constructed in several ways —such as gaussian distance-based topological adjacency (Li et al. 2017) or self-adaptive data-driven adjacency (Wu, Pan, et al. 2019)— and denote proximity of nodes for the GNN.

5.2 Investigation of Physics-Informed Modeling Strategies (Addressing RQ1)

To determine the optimal method for embedding traffic flow theory into GNNs (RQ1), this project will implement and systematically evaluate two distinct architectural strategies:

1. **Strategy A: Physics as Regularization (Soft Constraint):** This approach employs a standard Spatio-Temporal GNN backbone, consisting of a **Graph Attention Network (GAT)** to capture spatial dependencies followed by a temporal module (e.g. GRU or Transformer) to model time-series evolution. Physical consistency is implemented purely through a composite loss function

L_{total} (Di et al. 2023):

$$L_{total} = L_{data} + \lambda_{phy} L_{physics} \tag{1}$$

Here, $L_{physics}$ primarily penalizes predictions that violate the local Fundamental Diagram (FD) at the link level (e.g., ensuring flow does not exceed density-dependent capacity). Additionally, the Macroscopic Fundamental Diagram (MFD) is explored as a global regularization term to ensure that the aggregated network state remains within physically plausible bounds (Di et al. 2023).

2. **Strategy B: Physics as Architectural Embedding (Hard/Structure Constraint):** Building on the foundational framework of Physics-Informed Neural Networks (PINNs) (Raissi et al. 2019), this approach embeds physical laws directly into the model structure (Wang et al. 2024). It proposes a layer of "Local Fundamental Diagrams," where the neural network explicitly learns the parameters of the FD for each link. Crucially, as urban traffic is governed by signalized intersections, the learned "capacity" parameter will function as a proxy for the effective downstream throughput (saturation flow rate $\times$ effective green time), rather than solely the physical road geometry, aligning with the intersection-based capacity modeling approaches discussed in Fourati et al. (2021). These local physical relations then pass their outputs through the GAT-GRU structure. This forces the model to learn representations that are inherently grounded in local traffic physics before aggregating them spatially, potentially offering higher robustness in data-sparse regimes.

5.3 Latent Penetration Rate Module and Virtual Station Estimation (Addressing RQ3)

The final component addresses the challenge of estimating total flow on unobserved links ("Virtual Stations").

- **Latent Variable Learning:** Instead of assuming a fixed relationship between FCD and total flow, the architecture includes a dedicated sub-module that learns dynamic latent attributes, encapsulating the hidden structural and temporal complexities that bridge the sampled trajectories to the ground-truth volume. This approach draws on the Latent Space Model (LSM) framework of Deng et al. (2016), which demonstrates that traffic states on unobserved links can be inferred by learning dynamic latent attributes that capture topological and temporal dependencies. This module takes the *contextual features* extracted in the data pipeline (Section 5.1) as input and learns a non-linear mapping to the observed PR at sensor locations.
- **Network-Wide Estimation:** Once trained, this module infers the latent PR for the entire graph based on the ubiquitous contextual features available at every link (e.g., road type, time, POI density). By generalizing the latent attributes to edges without sensors—similar to the missing-sensor completion strategy in Deng et al. (2016)—the model generates total flow estimates for "Virtual Stations," effectively recovering the population flow from the biased sample across the entire unobserved network.

5.4 Evaluation: Inductive Spatial Cross-Validation

To validate the model's ability to estimate flows at "Virtual Stations" (RQ1), standard training-validation split methodology from machine learning is insufficient as it is inherently *transductive*—allowing the model to memorize location-specific biases rather than learning generalizable spatial dependencies. Instead, this research follows the inductive learning paradigm established by Wu, Zhuang, et al. (2021).

The set of graph nodes is partitioned into disjoint spatial subsets: V_{train} (observed) and V_{test} (unobserved). Following the protocol in Wu, Zhuang, et al. (2021), we randomly mask a percentage of nodes (e.g., 25%) during training to simulate unequipped segments. Evaluation is performed strictly on these withheld nodes, ensuring the model generalizes the message-passing mechanism to unseen network topologies rather than simply forecasting temporal trends for known sensors.

6 Scientific Contribution

The specific contributions are threefold:

1. **Methodological Integration of Physics and Deep Learning:** This research develops a Graph Neural Network architecture that explicitly integrates traffic flow theory. It fuses local constraints based on the Fundamental Diagram (FD) with global consistency checks potentially derived from the Macroscopic Fundamental Diagram (MFD). By constraining the loss function with physical Conservation Laws, the framework addresses the "black box" limitation of standard deep learning. This contributes a method for imputing traffic states in data-sparse regions where purely statistical models fail to generalize physically plausible results.

2. **Empirical Characterization of Penetration Rate Dynamics:** This work provides a comprehensive quantitative analysis of the spatiotemporal dynamics of commercial FCD Penetration Rates (PR) in an urban environment. Contrary to the common assumption of a static upscaling factor, the study identifies specific, non-linear determinants of PR variability.

3. **Scalable Network-Wide Estimation (The "Virtual Station"):** The project validates a workflow for estimating total traffic flow on street segments that lack physical sensors. By capturing the complex relationship between partial vehicle observations and network-wide flow as a latent variable conditioned on ubiquitous contextual features (e.g., road class, time), the research demonstrates the feasibility of software-based estimation as a complement to physical infrastructure. This directly addresses the economic and operational constraints of urban traffic monitoring by reducing the dependency on dense hardware networks (Ashwini and Sumathi 2020).

7 Work Plan

Formalities: Cumulative Dissertation (3+ papers) written in English.

Resource Planning: The project relies on historical Floating Car Data (FCD) provided by INRIX (secured via the EU-Horizon project *NEXTLOGIC*) and stationary detector data from Frankfurt am Main (accessible via the Mobilithek real-time API). Access to both datasets is pre-validated and secured. All Extract-Transform-Load (ETL) processes and computational experiments will be conducted on the AI virtual machines and high-performance GPU cluster provided by the Frankfurt University of Applied Sciences. The Ph.D. timeline is fully funded until the end of 2028 by the EU-Horizon project NextLogic.

Timeline and Milestones: The project is structured into three overlapping phases over a 36-month period, designed to incrementally address the research questions. Phase 1 has been completed (with the first manuscript currently under review at the European Transport Research Review), and I am currently in the second month of Phase 2.

Phase	Timeframe	Activity/Milestone
1	Months 1-6	**Data Pipeline & Empirical Analysis:** Implementation of the ETL pipeline for fusing INRIX FCD and loop detector data. Conducting the large-scale statistical analysis of Penetration Rate dynamics and determinants. *Milestone: Submission of Paper 1 (Dynamics of Commercial FCD Penetration Rates).*
2	Months 7-18	**Methodology Development (Imputation):** Development of the PI-GNN architecture. Implementation and comparative testing of physics-informed strategies (Loss vs. Architecture) to address data sparsity. *Milestone: Submission of Paper 2 (Physics-Informed Imputation Strategies).*
3	Months 19-30	**Model Extension (Network Estimation):** Extending the model to estimate flow at "virtual stations" by implementing the latent Penetration Rate learning module. Evaluation against held-out test sets. *Milestone: Submission of Paper 3 (Network-Wide Flow Estimation).*
4	Months 31-36	**Thesis Compilation:** Synthesis of published papers, writing the framing text (the introductory and concluding chapters)

Table 1: Project Timeline and Milestones

8 References

Afandizadeh, Shahriar, Saeid Abdolahi, and Hamid Mirzahossein (Jan. 2024). "Deep Learning Algorithms for Traffic Forecasting: A Comprehensive Review and Comparison with Classical Ones". In: *Journal of Advanced Transportation* 2024.1. Ed. by Ruimin Li, p. 9981657. ISSN: 0197-6729, 2042-3195. DOI: 10.1155/2024/9981657. URL: https://onlinelibrary.wiley.com/doi/10.1155/2024/9981657 (visited on 02/03/2026).

Ahmed, Mohammed S. and Allen R. Cook (1979). "ANALYSIS OF FREEWAY TRAFFIC TIME-SERIES DATA BY USING BOX-JENKINS TECHNIQUES". In: *Transportation Research Record* 722. ISSN: 0361-1981. URL: https://trid.trb.org/View/148123 (visited on 02/03/2026).

Ashwini, B. P. and R. Sumathi (Dec. 3, 2020). "Data Sources for Urban Traffic Prediction: A Review on Classification, Comparison and Technologies". In: *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*. 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS). Thoothukudi, India: IEEE, pp. 628–635. ISBN: 978-1-7281-7089-3. DOI: 10.1109/ICISS49785.2020.9316096. URL: https://ieeexplore.ieee.org/document/9316096/ (visited on 02/03/2026).

Chen, Xinyu, Zhaocheng He, and Lijun Sun (Jan. 2019). "A Bayesian tensor decomposition approach for spatiotemporal traffic data imputation". In: *Transportation Research Part C: Emerging Technologies* 98, pp. 73–84. ISSN: 0968090X. DOI: 10.1016/j.trc.2018.11.003. URL: https://linkinghub.elsevier.com/retrieve/pii/S0968090X1830799X (visited on 02/03/2026).

Cui, Zhiyong et al. (Aug. 2020). "Graph Markov network for traffic forecasting with missing data". In: *Transportation Research Part C: Emerging Technologies* 117, p. 102671. ISSN: 0968090X. DOI: 10.1016/j.trc.2020.102671. URL: https://linkinghub.elsevier.com/retrieve/pii/S0968090X20305866 (visited on 02/13/2026).

Daganzo, Carlos F. (Aug. 1994). "The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory". In: *Transportation Research Part B: Methodological* 28.4, pp. 269–287. ISSN: 01912615. DOI: 10.1016/0191-2615(94)90002-7. URL: https://linkinghub.elsevier.com/retrieve/pii/0191261594900027 (visited on 02/03/2026).

Deng, Dingxiong et al. (Aug. 13, 2016). "Latent Space Model for Road Networks to Predict Time-Varying Traffic". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco California USA: ACM, pp. 1525–1534. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939860. URL: https://dl.acm.org/doi/10.1145/2939672.2939860 (visited on 02/04/2026).

Di, Xuan et al. (June 17, 2023). "Physics-Informed Deep Learning for Traffic State Estimation: A Survey and the Outlook". In: *Algorithms* 16.6, p. 305. ISSN: 1999-4893. DOI: 10.3390/a16060305. URL: https://www.mdpi.com/1999-4893/16/6/305 (visited on 02/04/2026).

Fourati, Walid, Hekmat Dabbas, and Bernhard Friedrich (2021). "Estimation of Penetration Rates of Floating Car Data at Signalized Intersections". In: *Transportation Research Procedia* 52, pp. 228–235. DOI: 10.1016/j.trpro.2021.01.026.

Geroliminis, Nikolas and Carlos F. Daganzo (Nov. 2008). "Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings". In: *Transportation Research Part B: Methodolog-*

ical 42.9, pp. 759–770. ISSN: 01912615. DOI: 10.1016/j.trb.2008.02.002. URL: https://linkinghub.elsevier.com/retrieve/pii/S0191261508000180 (visited on 02/13/2026).

Greenshields, Bruce D. (1935). "A Study of Traffic Capacity". In: 14, pp. 448–477. URL: https://onlinepubs.trb.org/Onlinepubs/hrbproceedings/14/14P1-023.pdf (visited on 02/03/2026).

Guo, Shengnan et al. (July 17, 2019). "Attention Based Spatial-Temporal Graph Convolutional Networks for Traffic Flow Forecasting". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33.1, pp. 922–929. ISSN: 2374-3468, 2159-5399. DOI: 10.1609/aaai.v33i01.3301922. URL: https://ojs.aaai.org/index.php/AAAI/article/view/3881 (visited on 02/13/2026).

Hamann, Jonas and Tobias Hagen (Mar. 2025). "Revealing Trip Purposes in Raw GPS Data by Applying a Multi-Phase Clustering Approach to Semantic Trajectories". In: *IEEE Transactions on Intelligent Transportation Systems* 26.3, pp. 3543–3556. ISSN: 1524-9050, 1558-0016. DOI: 10.1109/TITS.2024.3516141. URL: https://ieeexplore.ieee.org/document/10811755/ (visited on 02/03/2026).

Jiang, Weiwei and Jiayun Luo (2022). "Graph Neural Network for Traffic Forecasting: A Survey". In: *Expert Systems with Applications* 207, p. 117921. DOI: 10.1016/j.eswa.2022.117921.

Kipf, Thomas N. and Max Welling (2016). *Semi-Supervised Classification with Graph Convolutional Networks*. Version Number: 4. DOI: 10.48550/ARXIV.1609.02907. URL: https://arxiv.org/abs/1609.02907 (visited on 02/13/2026).

Li, Yaguang et al. (2017). *Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting*. DOI: 10.48550/ARXIV.1707.01926. URL: https://arxiv.org/abs/1707.01926 (visited on 02/03/2026).

Lighthill, Michael James and Gerald Beresford Whitham (May 10, 1955). "On kinematic waves II. A theory of traffic flow on long crowded roads". In: *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 229.1178, pp. 317–345. ISSN: 0080-4630, 2053-9169. DOI: 10.1098/rspa.1955.0089. URL: https://royalsocietypublishing.org/rspa/article/229/1178/317/9548/On-kinematic-waves-II-A-theory-of-traffic-flow-on (visited on 02/03/2026).

Pan, Yuyan Annie et al. (June 2025). "Urban intersection traffic flow prediction: A physics-guided stepwise framework utilizing spatio-temporal graph neural network algorithms". In: *Multimodal Transportation* 4.2, p. 100207. ISSN: 27725863. DOI: 10.1016/j.multra.2025.100207. URL: https://linkinghub.elsevier.com/retrieve/pii/S2772586325000218 (visited on 02/03/2026).

Raissi, M., P. Perdikaris, and G.E. Karniadakis (Feb. 2019). "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations". In: *Journal of Computational Physics* 378, pp. 686–707. ISSN: 00219991. DOI: 10.1016/j.jcp.2018.10.045. URL: https://linkinghub.elsevier.com/retrieve/pii/S0021999118307125 (visited on 02/13/2026).

Richards, Paul I. (Feb. 1956). "Shock Waves on the Highway". In: *Operations Research* 4.1, pp. 42–51. ISSN: 0030-364X, 1526-5463. DOI: 10.1287/opre.4.1.42. URL: https://pubsonline.informs.org/doi/10.1287/opre.4.1.42 (visited on 02/03/2026).

Rizvi, Syed Muzammil Abbas and Bernhard Friedrich (Jan. 2024). "Optimizing Urban Traffic Management: Investigating the Impact of Probe Vehicle Road Coverage, Penetration Rates, and Speed

Data Quality in Macroscopic Fundamental Diagram Estimation". In: *Journal of Advanced Transportation* 2024.1. Ed. by Tao Liu, p. 2189950. ISSN: 0197-6729, 2042-3195. DOI: 10.1155/2024/2189950. URL: https://onlinelibrary.wiley.com/doi/10.1155/2024/2189950 (visited on 02/03/2026).

Saki, Siavash and Tobias Hagen (Aug. 4, 2022). "A Practical Guide to an Open-Source Map-Matching Approach for Big GPS Data". In: *SN Computer Science* 3.5, p. 415. ISSN: 2661-8907. DOI: 10.1007/s42979-022-01340-5. URL: https://link.springer.com/10.1007/s42979-022-01340-5 (visited on 02/03/2026).

Seo, Toru et al. (2017). "Traffic state estimation on highway: A comprehensive survey". In: *Annual Reviews in Control* 43, pp. 128–151. ISSN: 13675788. DOI: 10.1016/j.arcontrol.2017.03.005. URL: https://linkinghub.elsevier.com/retrieve/pii/S1367578817300226 (visited on 02/13/2026).

Treiber, Martin and Arne Kesting (2013). *Traffic Flow Dynamics: Data, Models and Simulation*. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN: 978-3-642-32460-4. DOI: 10.1007/978-3-642-32460-4. URL: https://link.springer.com/10.1007/978-3-642-32460-4 (visited on 02/13/2026).

Turksma, S. (2000). "The various uses of floating car data". In: *Tenth International Conference on Road Transport Information and Control*. Tenth International Conference on Road Transport Information and Control. Vol. 2000. London, UK: IEE, pp. 51–55. ISBN: 978-0-85296-725-6. DOI: 10.1049/cp:20000103. URL: https://digital-library.theiet.org/content/conferences/10.1049/cp_20000103 (visited on 02/03/2026).

Veličković, Petar et al. (2017). *Graph Attention Networks*. DOI: 10.48550/ARXIV.1710.10903. URL: https://arxiv.org/abs/1710.10903 (visited on 02/03/2026).

Wang, Ting et al. (Dec. 2024). "PI-STGnet: Physics-integrated spatiotemporal graph neural network with fundamental diagram learner for highway traffic flow prediction". In: *Expert Systems with Applications* 258, p. 125144. ISSN: 09574174. DOI: 10.1016/j.eswa.2024.125144. URL: https://linkinghub.elsevier.com/retrieve/pii/S0957417424020116 (visited on 02/03/2026).

Wu, Yuankai, Dingyi Zhuang, et al. (May 18, 2021). "Inductive Graph Neural Networks for Spatiotemporal Kriging". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35.5, pp. 4478–4485. ISSN: 2374-3468, 2159-5399. DOI: 10.1609/aaai.v35i5.16575. URL: https://ojs.aaai.org/index.php/AAAI/article/view/16575 (visited on 02/13/2026).

Wu, Zonghan, Shirui Pan, et al. (Aug. 2019). "Graph WaveNet for Deep Spatial-Temporal Graph Modeling". In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*. Twenty-Eighth International Joint Conference on Artificial Intelligence {IJCAI-19}. Macao, China: International Joint Conferences on Artificial Intelligence Organization, pp. 1907–1913. ISBN: 978-0-9992411-4-1. DOI: 10.24963/ijcai.2019/264. URL: https://www.ijcai.org/proceedings/2019/264 (visited on 02/03/2026).

Yu, Bing, Haoteng Yin, and Zhanxing Zhu (July 2018). "Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting". In: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*. Twenty-Seventh International Joint Conference on Artificial Intelligence {IJCAI-18}. Stockholm, Sweden: International Joint Conferences on Artificial Intelligence Organization, pp. 3634–3640. ISBN: 978-0-9992411-2-7. DOI: 10.24963/

ijcai . 2018 / 505. URL: https : / / www . ijcai . org / proceedings / 2018 / 505 (visited on 02/03/2026).

Zhan, Xianyuan et al. (Aug. 2013). "Urban link travel time estimation using large-scale taxi data with partial information". In: *Transportation Research Part C: Emerging Technologies* 33, pp. 37–49. ISSN: 0968090X. DOI: 10.1016/j.trc.2013.04.001. URL: https://linkinghub.elsevier.com/retrieve/pii/S0968090X13000740 (visited on 02/13/2026).

Zhang, Yimei et al. (Dec. 2024). "A Comprehensive Survey on Traffic Missing Data Imputation". In: *IEEE Transactions on Intelligent Transportation Systems* 25.12, pp. 19252–19275. ISSN: 1524-9050, 1558-0016. DOI: 10 . 1109 / TITS . 2024 . 3478816. URL: https : / / ieeexplore . ieee . org / document/10733755/ (visited on 02/03/2026).

9 Appendix