Chair of Transportation Systems Engineering
TUM School of Engineering and Design
Technical University of Munich

Heterogeneous Data Integration and Imputation for Large-Scale Transportation Simulation Calibration

Master's Thesis

M.Tech. Barun Das

July 31, 2026

Description

Calibrating large-scale agent-based transport simulation models requires accurate, consistent, and complete data from heterogeneous sources. In practice, these sources differ in spatial resolution, temporal granularity, and format, such as loop detector counts, travel survey data, GTFS feeds, floating car data, OpenStreetMap network attributes. No standardized pipeline currently exists that ingests, consolidates, quality-checks, and imputes these multi-source data streams into a unified representation suitable for automated calibration across multiple simulation scenarios.

Objective

This thesis will design and implement a reproducible data integration, ingestion and imputation pipeline for heterogeneous transportation data, applicable to at least three study cities (Dresden, Berlin, Munich). The student will define a unified spatio-temporal data representation, implement data connectors for key data sources, design and apply a data integrity assessment framework (missing value patterns, outlier detection), and develop or benchmark imputation methods suitable for sparse multi-modal transportation data.

Core Tasks

- Audit and document the data landscape across study cities: sources available, resolution, access conditions, and structural discrepancies
- Design a unified spatio-temporal data schema capable of representing point, link, and area-level mobility observations across multiple modalities
- Implement data connectors for key data sources (e.g., BASt, SrV survey data, GTFS, OSM attributes, floating car data)
- Develop a data integrity assessment framework: quantify missing value patterns (temporal, spatial, systematic), detect outliers, and classify severity of integrity issues per source
- Implement and compare imputation methods for sparse transportation data - from matrix factorization and low-rank tensor completion to learned spatio-temporal imputation models and evaluate imputation accuracy against held-out ground truth
- Produce a cross-city data quality report: which sources, in which cities, are sufficient for calibration; where imputation is unavoidable; and quantify the estimated uncertainty propagation into calibration objectives

Chair of Transportation Systems Engineering
TUM School of Engineering and Design
Technical University of Munich

Requirements

- **Strong programming skills** in Python; comfortable with data engineering, machine learning, pandas/numpy, and geospatial tools (geopandas, shapely)
- **Solid mathematical background** in statistics, linear algebra, and ideally tensor methods or matrix factorization
- **Familiarity with transportation data sources** (loop detectors, GTFS, OpenStreetMap, travel surveys) is a strong plus
- **Interest in simulation and calibration** to understand why data quality matters for model outputs

Starting Date

As soon as possible.

How to Apply

If interested, please email your resume, academic transcripts, and preferred start date to Barun Das (barun.das@tum.de). A brief cover letter explaining your interest in the topic and any relevant experience with transportation data or data engineering is encouraged.