technische universität
dortmund

Master's Thesis

Missing spatial-temporal multimodal traffic data flow imputation and prediction

Ankita Sarkar

March 6, 2025

TU Dortmund University

Agenda

- Introduction
- Problem statement
- Literature Review
- Methodology
- Dataset and Feature Engineering
- Experiment Design
- Results and Discussion
- Conclusion and Future Work

Introduction

Introduction

- Traffic flow data is critical for urban planning and intelligent transportation.
- Missing data from sensor failures and transmission errors is a significant challenge.
- **Thesis Goal:**
  - To address missing data imputation in spatial-temporal multimodal traffic datasets.

Problem Statement

Formal Definition:

- Graph $G = (V, E)$ where V = nodes, E = edges:
- Neighnourhood of a node v : $N_v(u) = \{u \in V \mid (v, u) \in E\}$
- Adjasecency matrix

$$A_{ij} = \begin{cases} 1, & \text{if } e_{ij} \in E, \\ 0, & \text{otherwise.} \end{cases}$$

- Node representation: $X \in \mathbb{R}^{V \times F}$
- Spatio-temporal graph: $G(t) = (V, E, X_t)$

Problem Statement

Thesis goal

- **Learn prediction function:** $f : \{G(t-T), G(t-T+1), \ldots, G(t)\} \rightarrow Y_{t+\tau}$ with T as timestamp and $\tau = 15min$.
- Input data and labels are sparse.
- **Geospatial Influence on Traffic Patterns:** Building footprints, water bodies, POIs.
- **Congestion Classes:** Green (uncongested), Yellow (moderate), Red (heavy).

Literature Review

Literature Review

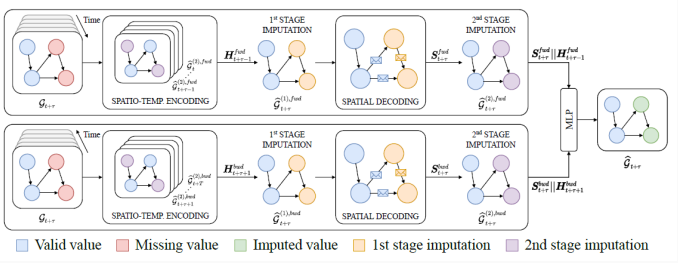
- **Classical Imputation Models:**
  - **Temporal:** HA, ARIMA, PPCA.
  - **Spatial:** Regression, Kriging.
  - **Spatial-Temporal:** Kernel PPCA, KNN.
- **Deep Learning Models:** LSTM, RNN, CNN, GAN.
- **Graph Neural Networks (GNNs):** GCN, STGCN, DSTGCN, GAT.
- **Limitations of Existing Models:** Incorporation of multimodality features.

Methodology

Methodology

- **Graph Recurrent Imputation Network (GRIN):**
  - Integrates Graph Neural Networks (GNNs) and Recurrent Neural Networks (RNNs).
  - Uses GCNs for spatial relations, GRUs for time-series imputation.
  - Imputates in both forward and backward directions.
- **Encoder Architecture:** Message Passing Graph Recurrent Network (MPNN + GRU).
- **Decoder Architecture:** Spatio-Temporal Graph Convolution.

Graph recurrent imputation network (Cini et al. [2021])



- Sequence of graphs G_t with nodes N_t at time step t.

$$G_t =< X_t, W_t >$$

- Binary mask for the missing values. M_t $\in \{0,1\} N_t \times d$

Dataset and Feature Engineering

Dataset

- **Dataset:**
  - Traffic4cast 2022 NeurIPS competition (London).
  - Loop counter data, building footprints.
  - Data Points:
    - Nodes: 59,110.
    - Edges: 132,414.
    - Loop Counters: 3,751.
    - Time Period: 2019-07-01 to 2020-01-31.
    - Sampling Rate: 15-minute intervals.
- **Temporal Features:** Speed, Vehicle counts.
- **Multimodal spatial features:** Building density, building diversity, building type encoding, geographic locations of loop counters and buildings.

Dataset

Road_graph

- Nodes - node_id, counter_info, num_assigned, longitude (x), latitude (y).
- Edges - node_id (start), node_id (end), speed_kph, parsed_maxspeed, importance, highway, oneway, lanes, tunnel.

speed_classes

node_id (start), node_id (end), day, t, volume_class, median_speed_kph, free_flow_kph.

	node_id	counter_info	num_assigned	x	y
0	78112			-0.145792	51.526976
1	99936			-0.152791	51.523611
2	99937			-0.152024	51.523018
3	101818	01/285	1	-0.148104	51.535179
4	101831	02/065	1	-0.147044	51.535612
...	...	...	...	...	...
59105	45951396121057865 18			-0.299336	51.588589
59106	82308311166816608 64			-0.037311	51.680737

	u	v	day	t	volume_class	median_speed_kph	free_flow_kph
0	78112	25508583	2020-01-31	29	3	40.941176	36.352941
1	78112	25508583	2020-01-31	30	5	10.823529	36.352941
2	78112	25508583	2020-01-31	31	5	41.647059	36.352941
3	78112	25508583	2020-01-31	32	5	22.901961	36.352941
4	78112	25508583	2020-01-31	34	5	48.000000	36.352941
...	...	...	...	...	...	...	...
3756047	48907014241332646 27	27596189	2020-01-31	86	5	22.352941	29.882353
3756048	48907014241332646 27	27596189	2020-01-31	87	1	35.764706	29.882353
3756049	48907014241332646 27	27596189	2020-01-31	89	1	29.647059	29.882353
3756050	48907014241332646 27	27596189	2020-01-31	90	5	21.411765	29.882353
3756051	48907014241332646 27	27596189	2020-01-31	92	5	25.647059	29.882353

Dataset

loop_counter

- index, node_id, day, counter_info, num_assigned, volume

	index	node_id	day	counter_info	num_assigned	volume
0	0	10028711	2019-07-01	[17/116]	[1]	[56.0, 44.0, 40.0, 31.0, 28.0, 22.0, 24.0, 16...
1	1	10028711	2019-07-02	[17/116]	[1]	[42.0, 35.0, 26.0, 21.0, 37.0, 34.0, 13.0, 20...
2	2	10028711	2019-07-03	[17/116]	[1]	[36.0, 23.0, 33.0, 41.0, 32.0, 30.0, 20.0, 9.0...
3	3	10028711	2019-07-04	[17/116]	[1]	[33.0, 32.0, 30.0, 18.0, 18.0, 14.0, 21.0, 18...
4	4	10028711	2019-07-05	[17/116]	[1]	[49.0, 40.0, 42.0, 41.0, 28.0, 19.0, 23.0, 18...
...	...	...	...	...	...	...
734801	107666	996609828	2020-01-27	[09/376]	[1]	[190.0, 188.0, 155.0, 120.0, 154.0, 144.0, nan...
734802	107667	996609828	2020-01-28	[09/376]	[1]	[157.0, 154.0, 137.0, 141.0, 114.0, 103.0, 82....

cc_labels

- node_id (start), node_id (end), day, t, cc

	u	v	day	t	cc
0	78112	25508583	2020-01-31	29	1
1	78112	25508583	2020-01-31	30	3
2	78112	25508583	2020-01-31	31	1
3	78112	25508583	2020-01-31	32	2
4	78112	25508583	2020-01-31	34	1
...	...	...	...	...	...
3756047	4890701424133264627	27596189	2020-01-31	86	2
3756048	4890701424133264627	27596189	2020-01-31	87	1

Dataset

osm_buildings

- osm_id, code, fclass, name, type, geometry

	osm_id	code	fclass	name	type	geometry
0	2956186	1500	building	Laurence House	block	POLYGON ((-0.02169 51.44459, -0.02168 51.44464...
1	2956187	1500	building	Lewisham Town Hall	None	POLYGON ((-0.02181 51.44498, -0.02161 51.44507...
2	2956188	1500	building	Broadway Theatre	None	POLYGON ((-0.02067 51.44542, -0.02064 51.44544...
3	2956192	1500	building	JD Sports	store	POLYGON ((-0.01903 51.44461, -0.01903 51.44462...
4	2956193	1500	building	Air Thrill	store	POLYGON ((-0.01834 51.44500, -0.01815 51.44551...
...	...	...	...	...	...	...
944003	1262637518	1500	building	None	school	POLYGON ((-0.36841 51.47726, -0.36839 51.47739...
944004	1262637520	1500	building	None	school	POLYGON ((-0.36930 51.47734, -0.36923 51.47775...

Exploratory Data Analysis

Exploratory Data Analysis: Congestion patterns, speed classes.

Feature Engineering

- Spatial features:
  - **Building density:** Number of buildings per unit area.
  - **Building type diversity:** Number of unique building types.
  - **Location type:** Residential, commercial, educational, etc.
- Why are these features relevant to traffic flow?
- **Temporal Features:** Time of day, day of week.
- **Multimodal Integration:** How are spatial and temporal features combined in the model?

Experiment Design

Experiment Design

- **Train-Validation-Test Split:**
  - Time series forecasting setup (Window-Horizon Approach).
  - 70% Training, 15% Validation, 15% Testing.
- **Mask Generation:** Training Mask, Evaluation Mask.
- Sequential Graph-Based Time Series Dataset Construction
- **Graph construction:**
  - Euclidean Distance Computation
  - Adjacency Matrix Calculation
  - Similarity Score: Gaussian Radial Basis Function - RBF
- **Evaluation Metrics:**
  - RMSE (Root Mean Squared Error).
  - MAE (Mean Absolute Error).
  - MAPE (Mean Absolute Percentage Error)
- **Model Parameters:** 8a38982 trainable parameters, 300 epochs, batch size 32.

Conclusion and Future Work

Conclusion and Future Work

- Explore different instantiations of GNN.
- Integrate additional modalities (weather, accidents, calendar holidays).
- Generalize to other cities and datasets.
- Integrate with traffic forecasting models - GAT.

References

Andrea Cini, Ivan Marisca, and Cesare Alippi. Filling the g_ap_s: Multivariate time series imputation by graph neural networks. *arXiv preprint arXiv:2108.00298*, 2021.

Thank you