8th IRTAD International Conference: Better Road Safety Data for Better Safety Performance
15-17 Athens 2026, Athens, Greece

Probabilistic Modeling for Node-Based Partitioning of Telematics-Informed Road Networks

Simone Paradiso[1,2], Apostolos Ziakopoulos[1], Petros Fortsakis[2], George Yannis[1]

[1] National Technical University of Athens, Department of Transportation Planning and Engineering, Athens, Greece
[2] OSeven Telematics, Chalandri, Greece

Introduction

Road Safety Overview

- **Road transport** is the most common mode of travel in the EU and crucial for the economy.
- The European Commission adopted Vision Zero to aim for zero road deaths and serious injuries by 2050.

Tech & Data for Road Safety

- IoT and AI are enhancing road safety by leveraging large-scale data.
- Data from sensors, cameras, and mobile devices enable pattern recognition in driving.
- Telematics data provide **vehicle trajectories** and **driving indicators** for deeper insights into driving behaviour.

Graph-Based Models

- Transformer-based **Graph Neural Networks** (GNNs) enable deep learning on graph-structured data.
- They can produce **node embeddings**: vector representations capturing global topology and neighborhood context.

Study Objectives

- The study explores driving behaviours mapped onto a **road network** using telematics data.
- Gaussian Mixture Model (GMM) is used for **probabilistic node-based clustering**.
- Transformer-based GNNs enhance **node representations before clustering**, improving insights into driving patterns.

Telematics Data Workflow

Data Collection

- Smartphone hardware sensors capture driver behaviour via app developed by **OSeven Telematics**.
- GDPR-compliant (Greek & EU regulations).

Data Preprocessing

- Proprietary machine learning algorithms process driving behavior data.
- Accuracy validated via:
  - OBD data
  - On-road tests
  - Literature benchmarks

Driving Trips Data

- 1,027 **anonymized trips** in Thessaloniki, Greece.
- Collected **Oct–Dec 2024**, at **1 Hz** frequency
- Contains per-second:
  - Trip coordinates & speed
  - Binary flags for
    - Harsh braking / acceleration
    - Speeding
    - Mobile use
    - Harsh event intensity: scale 1–3

Figure 1: *OSeven data flow system.*

Spatial Data Aggregation

Define Study Area

- Extract road network from **OpenStreetMap** for a specified bounding box, providing also associated road attributes.
  - Nodes = intersections or dead-ends.
  - Edges = road links between nodes.

Aggregate Telematics Data

- Edges: assign telematics observations to closest edge.
- Nodes: assign telematics observations within 50 m buffer, only if on connected edges to the considered node.

Table 1: *Telematics and Road Features for Nodes and Edges*

Feature	Applies To	Description
smoothenedSpeed	Nodes & Edges	Average speed near or along the spatial entity per trip
SpeedingFlag_per_trip	Nodes & Edges	Count of speeding events occurred near or along the spatial entity per trip
mobile_usage_per_trip	Nodes & Edges	Count of speeding events occurred near or along the spatial entity per trip
harsh_acc_per_trip	Nodes & Edges	Count of harsh acceleration events near or along the spatial entity per trip
harsh_brk_per_trip	Nodes & Edges	Count of harsh braking events near or along the spatial entity per trip
event_intensity	Nodes & Edges	Average intensity score of harsh events occurring near or along the spatial entity per trip
speed_std	Nodes & Edges	Speed standard deviation near or along the spatial entity per trip
street_count	Nodes Only	Number of streets connected to the node
edge_length	Edges Only	Length of the road
street_type	Edges Only	Categorical variable for street type (Urban, Rural, Service)
one_way	Edges Only	Direction of traffic (bidirectional or unidirectional)

Methods

This study integrates telematics data onto a road network to derive two node-level partitions based on driving behavior.

First, a **GMM** is applied to node features to estimate probabilistic behavior profiles, producing a behavior-based partition.

Next, a two-layer **Transformer-based GNN**, trained in a self-supervised framework, learns context-aware node representations, called node embeddings, by incorporating structural and edge information. These node embeddings are then used to generate a second partition by employed the GMM.

The **two partitions** are subsequently **compared**.

Clustering Comparison

Evaluation Metrics for K Clusters

- **Calinski–Harabasz Index** → measures cluster separation

$$CHI = \frac{Tr(B_k)/(K-1)}{Tr(W_k)/(N-K)}$$

- where B_k is between-cluster dispersion matrix, W_k = within-cluster dispersion matrix, N = total number of nodes.

- **Davies–Bouldin Index** → measures cluster compactness

$$DBI = \frac{1}{K}\sum_{i=1}^{K} \max_{j \neq i}\left(\frac{S_i + S_j}{M_{ij}}\right)$$

- where S_i = average distance of points in cluster i to its centroid and M_{ij} = distance between centroids of clusters i and j

Table 2: *Comparison of Clustering Performance*

K	Input	CHI (↑ = better between-cluster dispersion)	DBI (↓ = tighter, more distinct clusters)
2 (Selected by AIC)	Raw Features	1000.30	2.00
	Embeddings	901.98	1.68

Interpretation

Raw features → clusters more separated but less compact.

Node embeddings → clusters more compact with good separation.

Discussions

- Clustering on embeddings produces more meaningful and **well-separated clusters**, as indicated by a lower DBI. In contrast, raw features show higher between-cluster distances (CHI) but more poorly separated clusters.
- Using **embeddings**, the two clusters identified (Table 3) correspond to distinct driving behaviors:
  - **Cluster 1** refers to nodes where drivers tend to travel at lower speeds, rarely speeding and using their mobile phones, and showing no harsh maneuvers.
  - **Cluster 2** corresponds to nodes where drivers tend to have more hazardous driving behaviors, with higher average speeds, more frequent speeding, greater mobile usage and more variable driving patterns.
  - This framework allows for a **better understanding of driving patterns**, enabling targeted interventions that could improve road safety outcomes.

Table 3: *Average Feature Values Across Clusters*

Feature	Mean — Cluster 1 (3,643 Nodes)	Mean — Cluster 2 (2,663 Nodes)
smoothenedSpeed	22.28	35.45
SpeedingFlag_per_trip	0.01	0.25
mobile_usage_per_trip	0.14	0.46
harsh_acc_per_trip	0	0.01
harsh_brk_per_trip	0	0.01
event_intensity	0.08	0.28
speed_std	5.50	7.80
street_count	3.33	3.19

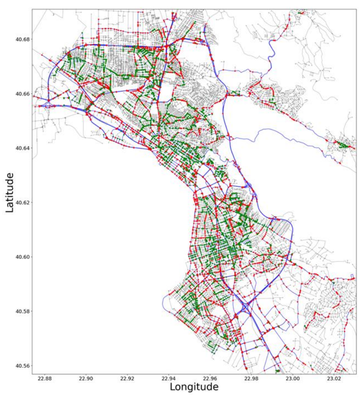

Figure 2: *Clustering of Thessaloniki Road Network Nodes Using Embedding*

Conclusions

- This GNN-derived embeddings used as input to the GMM model produced clusters that are more compact with lower intra-cluster distances.
- This enables the **identification of different groups of nodes**, for instance groups of nodes where people are more careful or where people tend to be more aggressive, facilitating **more coordinated and effective targeted interventions** aimed at improving road safety.
- While the study focuses on GMM models, **alternative models** could be used to explore different road network partitions.
- Examining **various GNN architectures** and experimenting with **different loss functions** may provide additional insights into graph partitioning.
- Including **additional features** and accounting for the **temporal dimension** could further improve the model's effectiveness in real-world applications.

Acknowledgments

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101119590

Contact Information:
Simone Paradiso, MSCA PhD Candidate and Researcher
Department of Transportation Planning and Engineering
Email: simone_paradiso@mail.ntua.gr
Website: https://linktr.ee/simone_paradiso