IRTAD Conference 2026
Better Road Safety Data for Better Safety Performance
15-17 April 2026 Athens, Greece
International Transport Forum
National Technical University of Athens

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

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

Department of Transportation Planning and Engineering, National Technical University of Athens, 5 Iroon Polytechneiou Street, 15773, Athens, Greece, simone_paradiso@mail.ntua.gr (Simone Paradiso)
Department of Transportation Planning and Engineering, National Technical University of Athens, 5 Iroon Polytechneiou Street, 15773, Athens, Greece, apziak@central.ntua.gr (Apostolos Ziakopoulos)
OSeven Telematics, 27B Chaimanta Street, 15234 Chalandri, Greece, pfortsakis@oseven.io (Petros Fortsakis)
Department of Transportation Planning and Engineering, National Technical University of Athens, 5 Iroon Polytechneiou Street, 15773, Athens, Greece, geyannis@central.ntua.gr (George Yannis)

Keywords: Spatial Road Safety, Telematics Data, Graph Neural Networks, Gaussian Mixture Models

Abstract

Background

Road transport is the most widely used means of travel in the European Union. While being essential to the economy in terms of its contribution to GDP, road transport is also the leading cause of crashes, serious injuries from collisions and premature deaths in Europe (European Climate, Infrastructure and Environment Executive Agency, 2022). The European Commission adopted the Vision Zero philosophy (Belin et al., 2012) with the goal of achieving zero road deaths and serious injuries by 2050.

Road safety has been boosted by the advancement of the Internet of Things (IoT) and Artificial Intelligence (AI) (Bhattacharya et al., 2022). The latter has been driven by the growing availability of data, which serve as training material for identifying and extracting patterns and that are increasingly collected through remote sensors, cameras, microphones, and mobile devices (Stylianou et al., 2019). Telematics data, which are a particularly informative data subset, have opened new research venues in road safety by providing vehicle trajectories and a variety of indicators (Ziakopoulos et al., 2022), allowing a deeper study of driving behaviour.

Furthermore, advancements in the field of AI have led to the development of the Transformer architecture, which has in turn been integrated into the field of Graph Neural Networks (GNNs), a subfield of deep learning that enables deep learning models field to deal with graph-structured data (Min et al., 2022). These graph-based architectures can be used to obtain a better node representation, known as node embedding, within the graph. The representations would be vectors of real numbers, incorporating topological relationships and neighbouring context, that can subsequently be used as inputs for other models to analyse the structure of the graphs and hence of a road network.

The present study aims to explore driving behaviours mapped onto a road network, therefore studying a telematics-informed network, using a probabilistic machine learning model. A Gaussian Mixture Model (GMM) was employed in order to obtain a probabilistic node-based clustering to explore the structure of telematics-informed nodes within a study area. The analysis also involved the use of Transformer-based GNN to enhance the quality of node representations used as input for the GMM.

Methods

Telematics data were provided by OSeven Telematics (OSeven, 2025), which collects driver behavior using smartphone hardware sensors, in compliance with Greek and European data protection regulations (GDPR). It utilizes Application Programming Interfaces (APIs) to retrieve sensor data, temporarily stored in the smartphone's database and then transmitted to the central back-end database (Kontaxi et al., 2023). The driving behavior data are pre-processed using machine learning algorithms, which cannot be disclosed due to intellectual property restrictions, whose accuracy has been validated against OBD data, on-road tests and literature benchmarks.

[1] * Corresponding author. Tel.: +30.2107721575;
E-mail address: simone_paradiso@mail.ntua.gr

The provided dataset consists of 1,027 anonymized naturalistic driving trips within the city of Thessaloniki in Greece, collected over the last four months of 2024 at a frequency of 1 Hz. The dataset contains trip coordinates along with per-second speed data, periodic binary flags (1/0) denoting the presence of harsh braking or harsh acceleration events, speeding and mobile use. It also includes the intensity of the harsh events on a scale from 1 to 3. In addition, a graph can be extracted by defining a bounding box from OpenStreetMap (OSM) (OpenStreetMap, 2025), a free and collaboratively edited map released under an open-content license. OSM provides detailed road network data which serves as the geometric framework for spatial analysis, alongside some features related to nodes and edges within the extracted graph. The configuration of the OSM query parameters determines the graph, where nodes correspond to "true" edge endpoints (i.e., intersections or dead-ends) and edges define the links between these nodes (Boeing, 2024).

As part of the preprocessing stage, telematics data were integrated to OSM entities to produce a telematics-informed road network. Telematics features were aggregated to the edges by using a closest-edge method where each telematics observation was first matched to its nearest edge, and then the matched data were aggregated per edge. The node aggregation was achieved by an enhanced buffer-based approach where telematics observations within a 50 m radius were assigned to a node only if located on edges connected to it.

Binary features (such as harsh braking, harsh acceleration, speeding flag and mobile phone usage) were aggregated by summing their occurrences for each corresponding spatial entity and then normalized by the number of trips at that entity to account for exposure. The 'event intensity' feature was averaged over occurrences of harsh braking or harsh acceleration per spatial entity, providing a measure of overall harsh-driving intensity at each node or edge. Finally, per-second speed measurements were averaged for each spatial entity, and their standard deviation was included in the dataset as a measure of overall traffic flow at the corresponding node or edge. The number of streets connected to each node, edge length, a manually encoded categorical variable for street type, and a binary flag indicating one-way streets were retained from OSM.

The present study aims to obtain a node-based partition of the road network. The telematics-informed nodes were first used as input for a Gaussian Mixture Model, a parametric probability density function represented as a weighted sum of Gaussian component densities (Reynolds, 2009). The model functions as a soft-clustering method, assigning each node a probability of belonging to one of K clusters, where K is a hyperparameter selected by researchers indicating K Gaussian distributions, each with distinct parameters capturing different driving behaviour patterns at node level. Each node is characterized by a vector of probabilities of belonging to these Gaussian distributions. Additionally, the present study also used a Transformer-based GNN (Shi et al., 2021), which is a type of GNN that uses transformer-style attention when combining data. The model was implemented following the detailed formulation documented in the PyTorch documentation (TransformerConv, 2025), and used for the node embedding task to generate richer node representations by considering information from all other nodes in the network and incorporating edge features.

Results and Discussions

Key Insights from GMM on Raw Features

The GMM was applied to the scaled raw features selecting $K = 2$, meaning that two different Gaussian distributions were used to model the underlying data. This value of K was selected, as it yielded the lowest AIC score while keeping the analysis interpretable. Subsequently, cluster labels were assigned based on the highest probability assigned to each sample within the node dataset. Results are presented in **Table 1**.

Table 1: Average Feature Values Across Clusters (Raw Features)

Feature	Mean — Cluster 1 (5,056 Nodes)	Mean — Cluster 2 (1,250 Nodes)
street_count	3.27	3.26
smoothenedSpeed	25.67	36.66
SpeedingFlag_per_trip	0	0.58
mobile_usage_per_trip	0.29	0.24
harsh_acc_per_trip	0	0.04
harsh_brk_per_trip	0	0.03
event_intensity	0	0.84
speed_std	5.49	10.44

The GMM produced two distinct clusters. The first one is a group of nodes where there no harsh events or speeding events occur on average These nodes are also characterized by lower average speeds but slightly higher mobile phone usage per trip. In contrast, the second cluster includes nodes associated with a higher

frequency of hazardous events. Figure 1 illustrates this pattern, where nodes belonging to the latter cluster are shown in red, and those in the former cluster are shown in green.

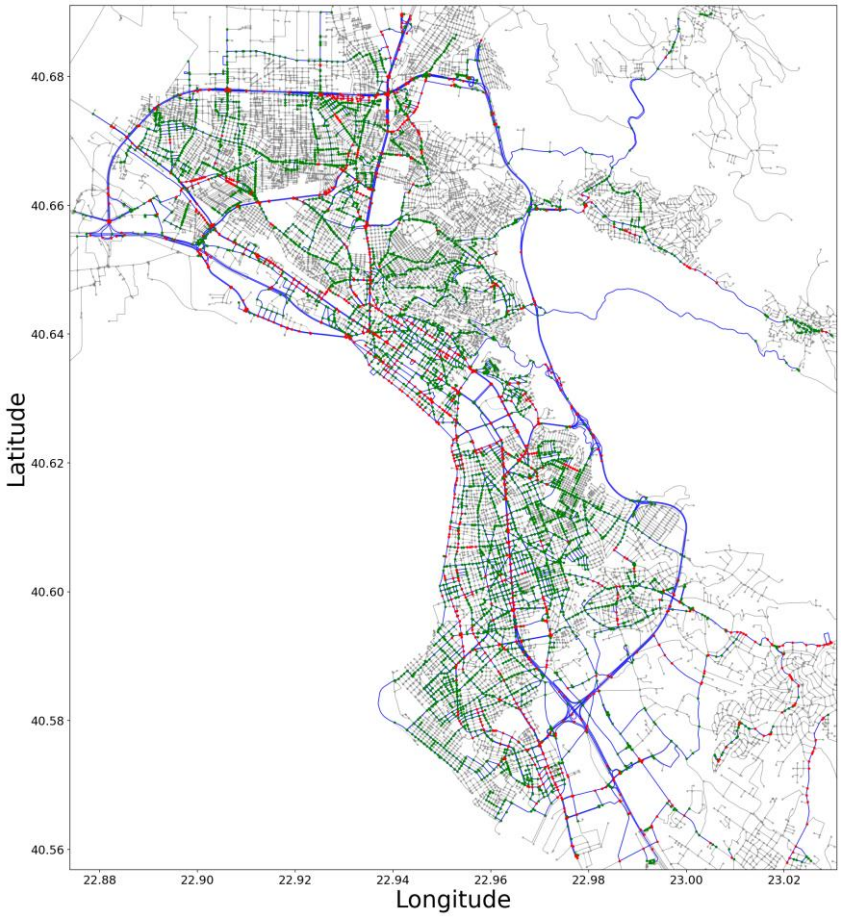

Figure 1: Clustering of Thessaloniki Road Network Nodes Using Raw Features

Key Insights from GMM on Feature Embeddings

A GNN with two Transformer-based layers was used to generate node embeddings that incorporate both neighbouring node features and edge features. The model was trained in a self-supervised way using a contrastive loss function designed by the authors, inspired by existing literature (Chen et al., 2020). The loss function creates an augmented graph by injecting Gaussian noise to the features, preserving the road network structure, while pulling each node closer to its augmented representation and pushing it away from all other nodes. The resulting embeddings were used as GMM input. $K = 2$ was again selected for reasons of interpretability and AIC score, as observed with the raw features. The obtained cluster labels were mapped back to the raw features to compute average feature values within each cluster. Results are shown in **Table 2**.

Table 2: Average Feature Values Across Clusters (Embeddings)

Feature	Mean — Cluster 1 (3,643 Nodes)	Mean — Cluster 2 (2,663 Nodes)
street_count	3.33	3.19
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

The partition appears to be enhanced as the first group exhibits lower values across all telematics features, likely representing a group of nodes where drivers behave more calmly. In contrast, the second group is characterized by higher telematics values, indicating drivers who are more aggressive, distracted, and prone

to harsh events. Furthermore, the traffic flow is more dispersed as suggested from the higher speed standard deviation.

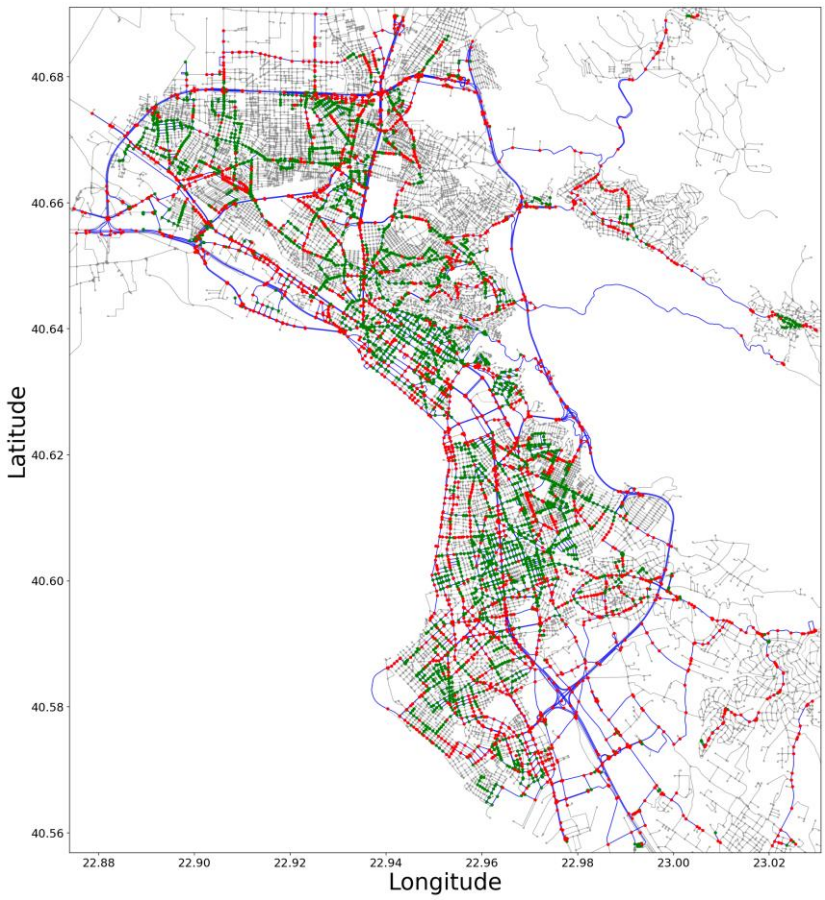

Figure 2: Clustering of Thessaloniki Road Network Nodes Using Embedding Features

To evaluate both cluster outputs, two metrics were used: the Calinski-Harabasz index (CHI) and the Davies-Bouldin index (DBI). The CHI is based on the ratio of between-cluster dispersion to within-cluster dispersion. The DBI evaluates clustering quality by combining cluster dissimilarity and intra-cluster dispersions (Hassan et al., 2021). The index values are presented in Table 3 below.

Table 3: Clustering Evaluation Metrics

K	Features	Calinski-Harabasz (CHI)	Davies-Bouldin (DBI)
2	Raw Features	1000.30	2.00
	Node Embeddings	901.98	1.68

The CHI favors the clustering based on raw features, whereas the DBI favors the embedding-based partition. This apparent contradictory result arises from the way the indices function: while the CHI suggests that raw features have larger between-cluster separation, the DBI suggests that node embeddings produce smaller average intra-cluster distance, more compact clusters and a reasonably large distance between cluster centroids. Although raw features can produce clusters that are more separated, they may not be as compact as those obtained from clustering on embeddings. Therefore, the analysis focused on the second approach yielding compact clusters that better reflect nodes that are closer to each other within the same cluster.

The work demonstrates how driving behavior data from telematics can be used to construct telematics-informed road networks, which serve as input for partitioning spatial entities using GMMs. Furthermore, the input to the GMM can be enhanced through GNN architectures, improving node representations and yielding clusters that are more compact and consistent than the those obtained through raw features.

The two clusters found with this approach represent distinct driving behaviors: one of them refers to nodes where drivers tend to travel at lower speeds, rarely speeding and using their mobile phones, and showing almost no harsh maneuvers. The other cluster corresponds to nodes where drivers tend to have more hazardous driving behaviors, with higher average speeds, frequent speeding, greater mobile usage and more variable

driving patterns. This framework allows for a better understanding of driving patterns and identifies areas where drivers are characterized by hazardous behavior, enabling targeted interventions that could improve road safety outcomes.

Conclusion

The current work proposes a novel approach for understanding telematics-informed road network data. After the spatial aggregation of telematics features, the study explored GMM to generate probabilistic soft-clustering labels, assigning each spatial entity a vector of probabilities representing its likelihood of belonging to each cluster. Specifically, this study focuses on a node-based partitioning of the road network, aiming to identify groups of nodes, i.e. intersections, characterized by drivers exhibiting similar driving behaviours, enabling targeted interventions for specific group of nodes. A Transformer-based GNN was applied to enhance the representation of the node capturing both topological structure and neighboring context. This enriched representation was then used as input to the GMM model, producing 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. It provides guidance on selecting the representations that leads to more bias-free results when focusing on the compactness of the targeted groups.

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.

References

Belin, M.-Å., Tillgren, P., & Vedung, E. (2012). Vision Zero – a road safety policy innovation. International Journal of Injury Control and Safety Promotion, 19(2), 171–179. https://doi.org/10.1080/17457300.2011.635213

Bhattacharya, S., Jha, H., & Nanda, R. P. (2022). Application of IoT and Artificial Intelligence in Road Safety. 2022 Interdisciplinary Research in Technology and Management (IRTM), 1–6. https://doi.org/10.1109/IRTM54583.2022.9791529

Boeing, G. (2024). Graph Simplification Solutions to the Street Intersection Miscount Problem (No. arXiv:2407.00258). arXiv. https://doi.org/10.48550/arXiv.2407.00258

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A Simple Framework for Contrastive Learning of Visual Representations (No. arXiv:2002.05709). arXiv. https://doi.org/10.48550/arXiv.2002.05709

European Climate, Infrastructure and Environment Executive Agency. (2022). EU Road Safety: Towards "Vision Zero" - European Commission. https://cinea.ec.europa.eu/publications/digital-publications/eu-road-safety-towards-vision-zero_en

Hassan, I. H., Abdullahi, M., & Yusuf, S. A. (2021). Analysis of Techniques for Selecting Appropriate Number of Clusters in k-means Clustering Algorithm.

Kontaxi, A., Tzoutzoulis, D.-M., Ziakopoulos, A., & Yannis, G. (2023). Exploring speeding behavior using naturalistic car driving data from smartphones. Journal of Traffic and Transportation Engineering (English Edition), 10(6), 1162–1173. https://doi.org/10.1016/j.jtte.2023.07.007

Min, E., Chen, R., Bian, Y., Xu, T., Zhao, K., Huang, W., Zhao, P., Huang, J., Ananiadou, S., & Rong, Y. (2022). Transformer for Graphs: An Overview from Architecture Perspective (No. arXiv:2202.08455). arXiv. https://doi.org/10.48550/arXiv.2202.08455

OpenStreetMap. (2025). About OpenStreetMap—OpenStreetMap Wiki. https://wiki.openstreetmap.org/wiki/About_OpenStreetMap

OSeven. (2025). Oseven.io. Oseven.Io. https://oseven.io/

Reynolds, D. (2009). Gaussian Mixture Models. In Encyclopedia of Biometrics (pp. 659–663). Springer, Boston, MA. https://doi.org/10.1007/978-0-387-73003-5_196

Shi, Y., Huang, Z., Feng, S., Zhong, H., Wang, W., & Sun, Y. (2021). Masked Label Prediction: Unified Message Passing Model for Semi-Supervised Classification (No. arXiv:2009.03509). arXiv. https://doi.org/10.48550/arXiv.2009.03509

Stylianou, K., Dimitriou, L., & Abdel-Aty, M. (2019). Chapter 12 - Big Data and Road Safety: A Comprehensive Review. In C. Antoniou, L. Dimitriou, & F. Pereira (Eds.), Mobility Patterns, Big

Data and Transport Analytics (pp. 297–343). Elsevier. https://doi.org/10.1016/B978-0-12-812970-8.00012-9

TransformerConv. (2025). torch_geometric.nn.conv.TransformerConv—Pytorch_geometric documentation. https://pytorch-geometric.readthedocs.io/en/2.5.2/generated/torch_geometric.nn.conv.TransformerConv.html

Ziakopoulos, A., Petraki, V., Kontaxi, A., & Yannis, G. (2022). The transformation of the insurance industry and road safety by driver safety behaviour telematics. Case Studies on Transport Policy, 10(4), 2271–2279. https://doi.org/10.1016/j.cstp.2022.10.011