The Use of Graph Neural Networks for Clustering in Road Safety Analysis

Simone Paradiso

IVORY PhD Candidate and Researcher

oseven

Together with:
George Yannis & Apostolos Ziakopoulos

Department of Transportation Planning and Engineering
National Technical University of Athens

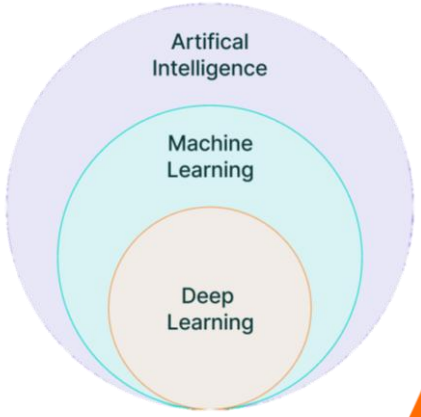
13th Symposium of the European Association for Research in Transportation (hEART2025)
10-12 June 2025, Munich, Germany

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

Introduction

- Road crashes claim 1.3 million lives annually, the leading cause of death for those under 29 and among the top 10 globally.

- IVORY framework.
  - European Union's Horizon Europe research and innovation programme Marie Skłodowska-Curie Industrial Doctorates (grant No 101119590).
  - It develops fair and explainable Artificial Intelligence (AI) to analyze driver behavior, predict crashes, and enhance road safety while sharing knowledge.
  - DC9 focuses on creating an AI framework to analyze road safety KPIs, predicts crashes, and evaluates the scalability of models primarily across spatial.



- Traditional crash prediction relies on econometrics; now enhanced by Machine Learning (ML) & Deep Learning (DL), with Graph Neural Networks (GNNs) extend DL to graph-structured data.

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

IVORY oseven

OSeven Telematics Data

- OSeven Telematics provided telematics data collected via smartphone hardware sensors to monitor driver behavior.
- All trips are anonymized and compliant with Greek and European personal data protection regulations (GDPR).
- Raw data are processed by proprietary machine learning algorithms. Reliability is validated against OBD data, on-road tests, simulators, and literature benchmarks.
- Selected features: geographic coordinates, smoothened speed, and binary flags for speeding, mobile usage, harsh acceleration, and harsh braking.

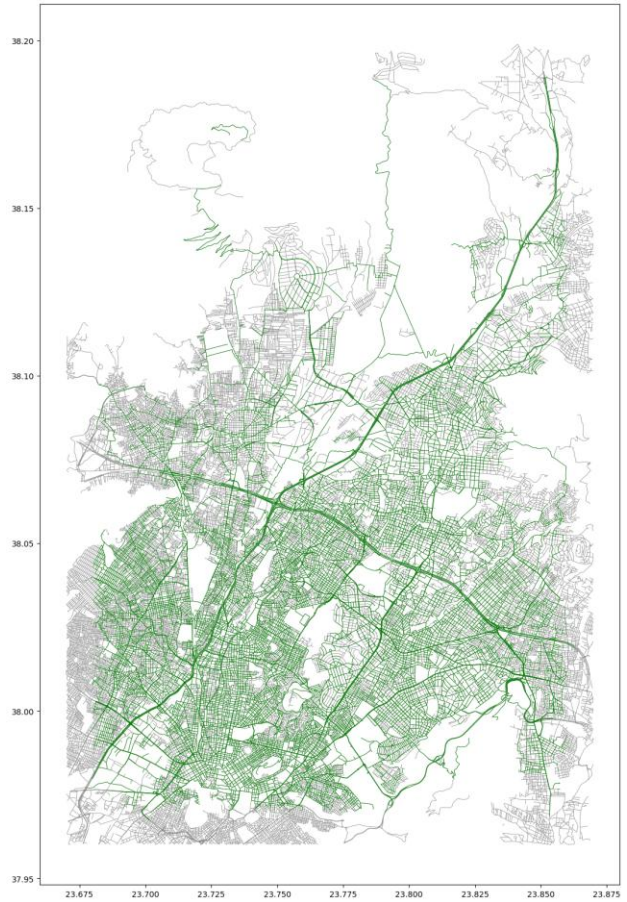
This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101119590

OpenStreetMap

- A free, editable global map created by volunteers and released under an open-content license.

- The previous dataset defines a coordinate bounding box used to extract a structured graph from OpenStreetMap via the OSMnx Python library.
  - Study area: urban road network in Athens, Greece.

- From the extracted graph, node and edge features were saved into two separate datasets.
  - These datasets were cleaned and preprocessed by removing irrelevant columns and ensuring data quality (e.g., handling duplicates and missing values).



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

Telematics Aggregation

- Telematics features were aggregated to OSM nodes using summation or averaging within a 50-meter buffer.
- Each raw telematics point was matched to its closest edge, and features were aggregated per edge similarly to the node-level process.
- The number of trips was calculated for each node and edge as exposure metric and the street_count attribute from OSM was added.
- Not all nodes/edges contain telematics data—coverage depends on trip paths.

Features	Description
Street_Count	Number of streets connected to the intersection
SmoothenedSpeed	Average speed of vehicles near the node
SpeedingFlag	Count of speeding events near the node
Mobile_usage	Number of instances of phone usage near the node
Harsh_acc	Number of harsh acceleration events near the node
Harsh_brk	Number of harsh braking events near the node
Trips_count	Number of trips recorded near the node

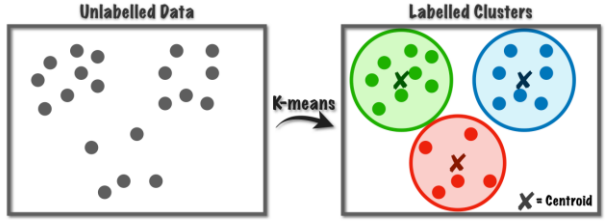
This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101119590

Clustering with K-means

- Given a dataset of points $\{x_1, x_2, \dots, x_n\}$, K-Means aims to partition them into K, clusters $\{C_1, C_2, \dots, C_k\}$, by minimizing the total within-cluster sum of squared distances:

$$arg\,\min_{C} \sum_{k=1}^{K} \sum_{x_i \in C_k} ||x_i - \mu_k||^2$$

Where μ_k is the centroid (mean) of cluster C_k, and $||x_i - \mu_k||^2$ is the squared Euclidean distance.



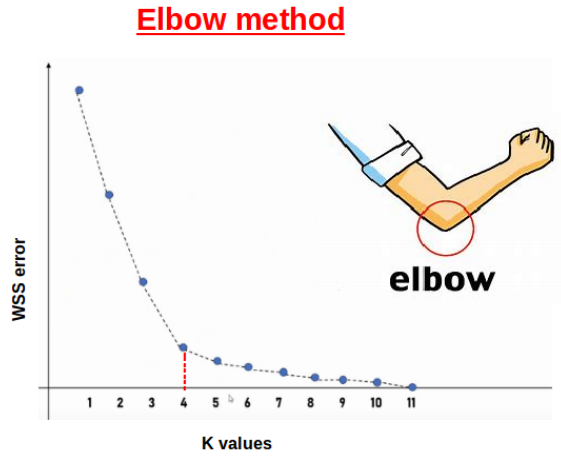
- K-Means: efficient and widely used.
  - Linear time complexity, fast convergence, low memory use.
  - Easy to implement and interpret, using clear centroids and simple assignments.
  - It is sensitive to initial centroids (can converge to local minima).

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

Choosing the Optimal Number of Clusters (K)

➢ Elbow Method

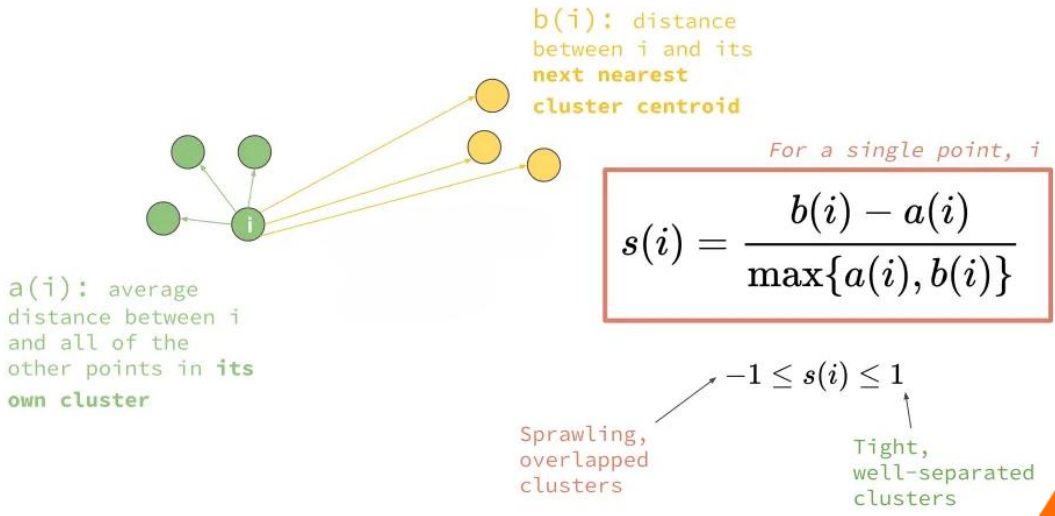
- Plots Within-Cluster Sum of Squares (WCSS) against K to identify the "elbow point" where adding more clusters no longer significantly reduces WCSS.
- This point indicates a good trade-off between model complexity and explained variance.



➢ Silhouette Score

- It measures how well a data point fits within its own cluster compared to other clusters.

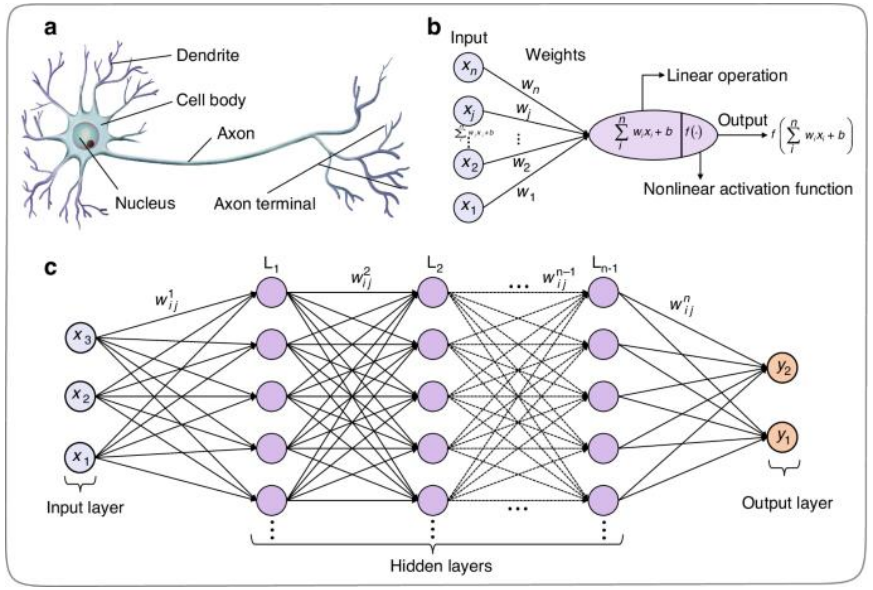
Average silhouette score quantifies overall cluster cohesion and separation.



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

Artificial Neural Networks

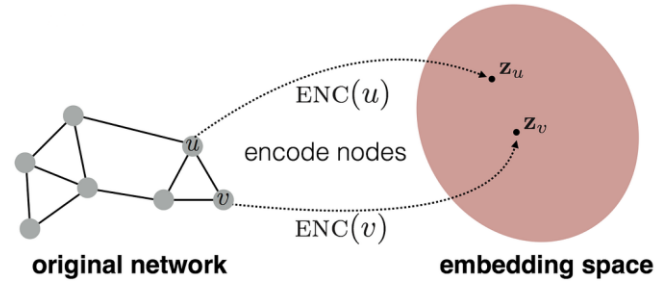
- Inspiration: Modeled after the structure of the human brain, consisting of layers of interconnected neurons.
- Structure:
  - Input layer receives raw data.
  - This data pass through one or multiple hidden layers that transform the input into data that is valuable for the output layer.
  - The output layer produces the predicted result based on the transformed information from the hidden layers.
- Key Concepts:
  - Activation functions introduce non-linearity.
  - Backpropagation updates parameters to minimize loss.
  - Loss Function guides learning, depending on the task.

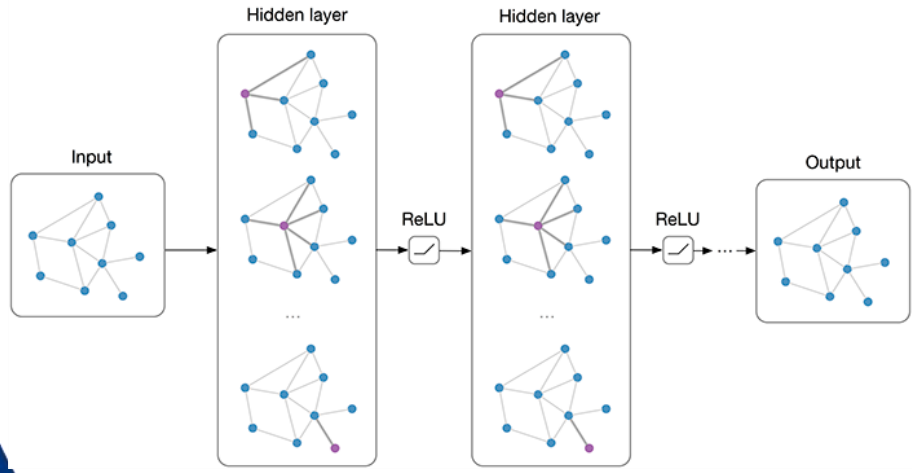


This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101119590

Introduction to GNNs

- GNNs are specialized artificial neural networks that are designed for tasks whose inputs are graphs.
- GNNs can be used to learn node embeddings: compact vector representations capturing each node's structural role, neighborhood context, and features.
- They encode graph-structured data by leveraging topological relationships rather than flattening the graph into vectors.
- Advancements include Graph Convolutional Networks (GCN) using convolution operations, and Graph Attention Networks (GAT) applying attention mechanisms.



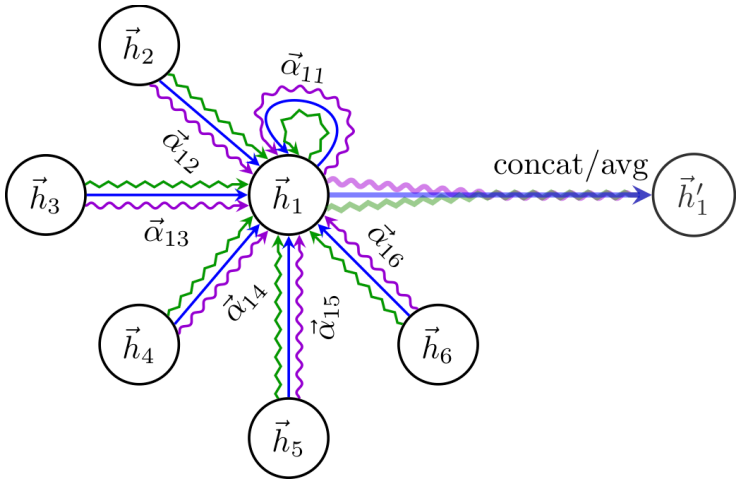


This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101119590

Attention-Based GNNs

- The GAT model is a novel neural network architecture leveraging attention mechanisms.
- It is a specialized form of GCN that improves how neighbor information is aggregated by learning attention weights dynamically.
- A simple neural network with two GAT layers has been defined.
  - Leverages attention coefficients (α_{ij}) to weigh neighbors differently and to incorporate both edge features and neighboring nodes.
  - Using a multi-head attention (${\alpha_{ij}}^{k}$) to stabilize training, improve accuracy and simplify output by averaging heads.

$$\boldsymbol{h}_i' = \sigma\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{j\in N_i\cup\{i\}} {\alpha_{ij}}^{k}\,\boldsymbol{W}^{\boldsymbol{k}}\boldsymbol{h}_{\boldsymbol{j}}\right)$$



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

Attention for Node Embeddings

- The attention coefficients quantify how much influence node j should have when updating node i's embedding.
  - A node is not equally influenced by all its neighbors, the GAT learns which surrounding nodes matter more prioritizing those that provide more meaningful context through attention.
- Edge Features Matter
  - If you include edge features, the model learns attention based on how nodes are connected, not just if they are connected.
- The attention coefficient α_{ij} can be interpreted as a learned weight of relevance:
  - High α_{ij} → Neighbor j is highly relevant for understanding node i's role in the network.
  - Low α_{ij} → Neighbor j is less informative for understanding node i's role in the network.
- Localized understanding of how risk or driver patterns propagate across a road network.

$$\alpha_{ij} = \frac{e^{(LeakyReLU(a_s^T \Theta_s x_i + a_t^T \Theta_t x_j + a_e^T \Theta_e e_{i,j}))}}{\sum_{k \in N_i \cup i} e^{(LeakyReLU(a_s^T \Theta_s x_i + a_t^T \Theta_t x_k + a_e^T \Theta_e e_{i,k}))}}$$

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

Self-Supervised Training

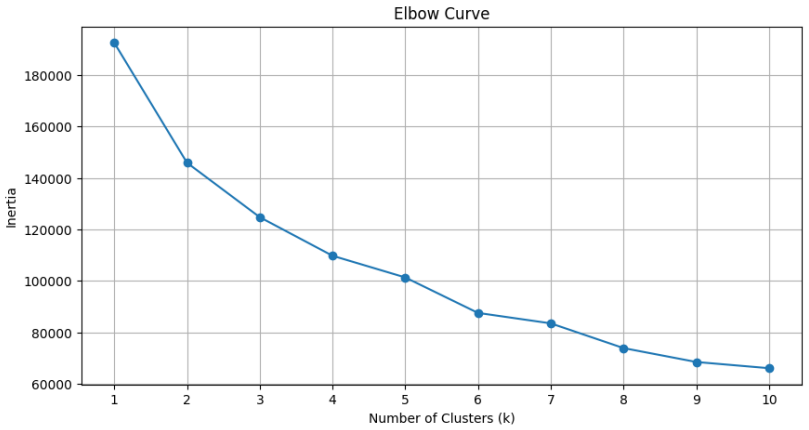
- Inspired by contrastive learning frameworks.
  - For each node, compute cosine similarity with neighbors and random non-neighbors.
  - Scale similarities with temperature and apply exponential function → $e^{S_i^+ j}$
  - Sum values over neighbors and non-neighbors separately.
  - Compute ratio of neighbor sum to total sum, then apply logarithm.
  - Final loss is averaged over all nodes, optimizing to learn embeddings that reflect shared characteristics and connectivity.
  - Training over 10 epochs using subgraphs sampled by PyG NeighborLoader.
  - Optimization done with Adam optimizer.

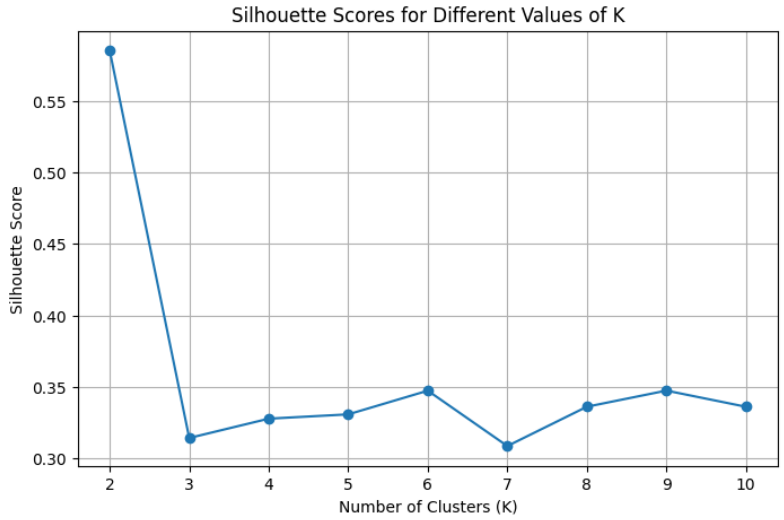
$$\mathcal{L}_i = -\log\left(\frac{\sum_{j\in P_i} e^{S_i^+ j}}{\sum_{j\in P_i} e^{S_i^+ j} + \sum_{j\in N_i} e^{S_i^- j}}\right)$$

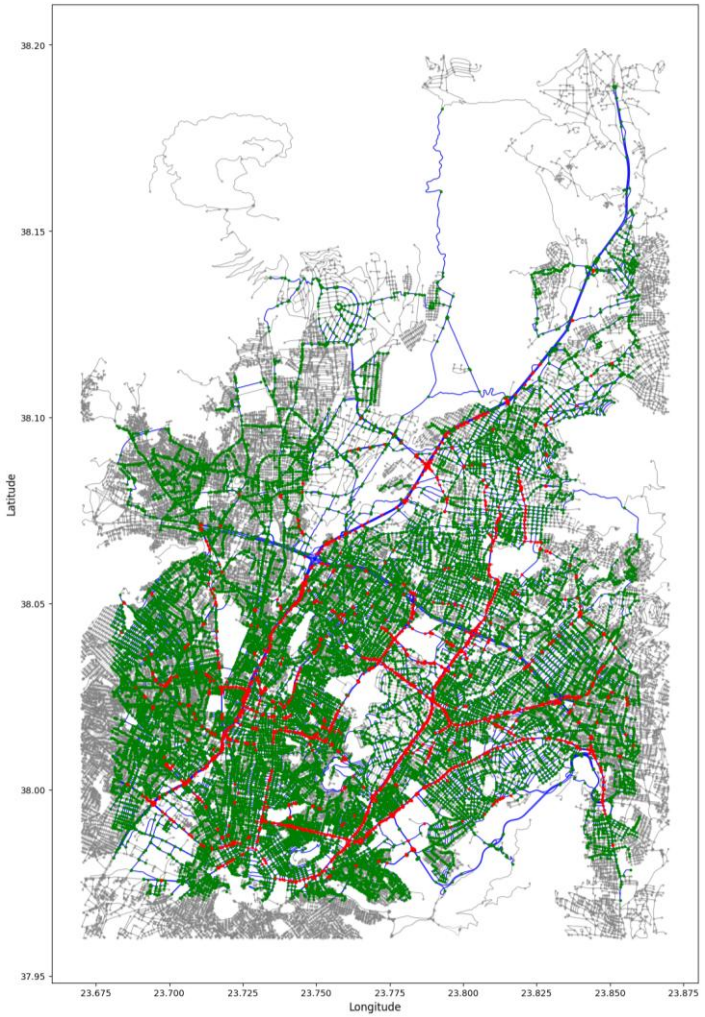
This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101119590

Clustering on Raw Features

- Clustering on Raw Features:
  - No clear elbow in the inertia curve → suggests gradual improvement without a sharp optimal K.
  - K-Means clustering on raw features yielded a highest silhouette score of 0.58 for K = 2→ modest separation between the two clusters.
  - Inertia of 145909, a 20% reduction from K=1 (inertia of 192556).





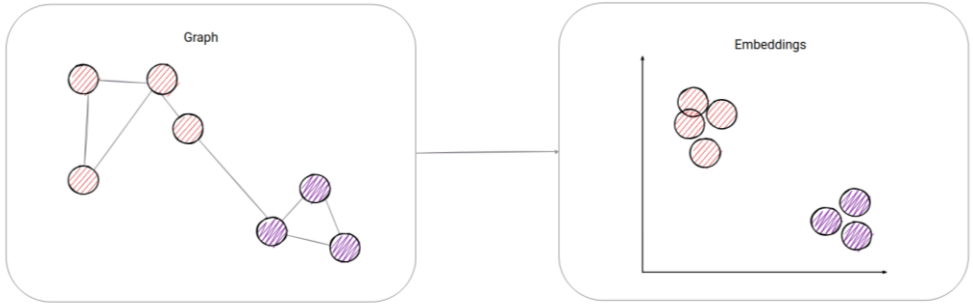


This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101119590

Embeddings

- At this stage, the GAT model was used to involve the features of the edges connected to each node, besides the road network topology.

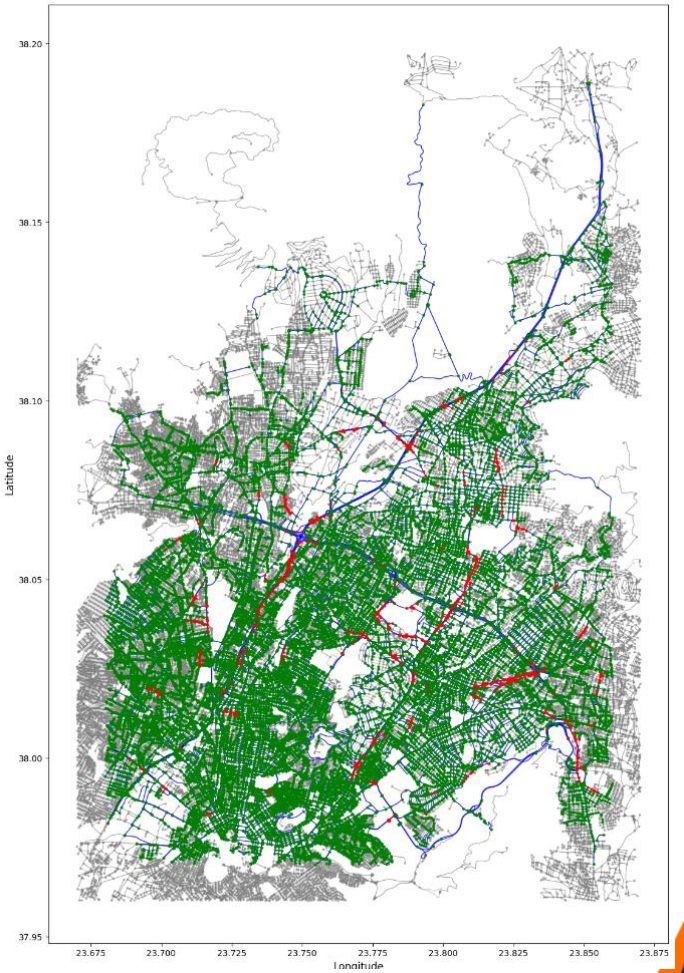
- The edge features mirror those of the nodes, excluding street_count and supplemented by four additional features:
  - Edge length.
  - Two binary features were derived via one-hot encoding from a three-category variable indicating road type (service, urban, rural).
  - A binary oneway column shows if vehicles can go in only one direction or in both directions.

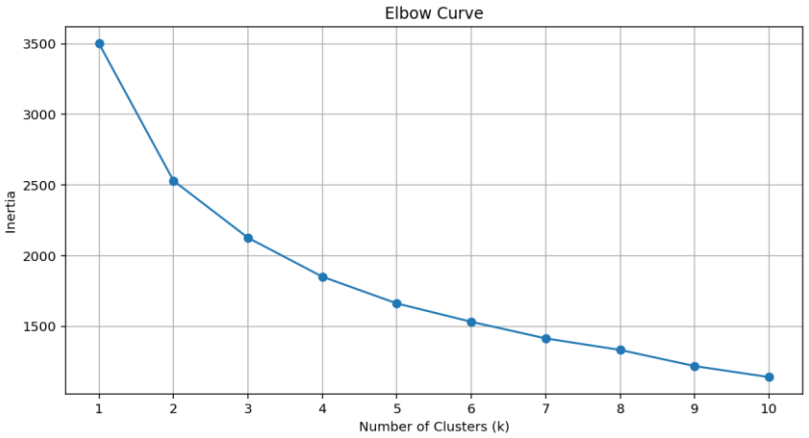


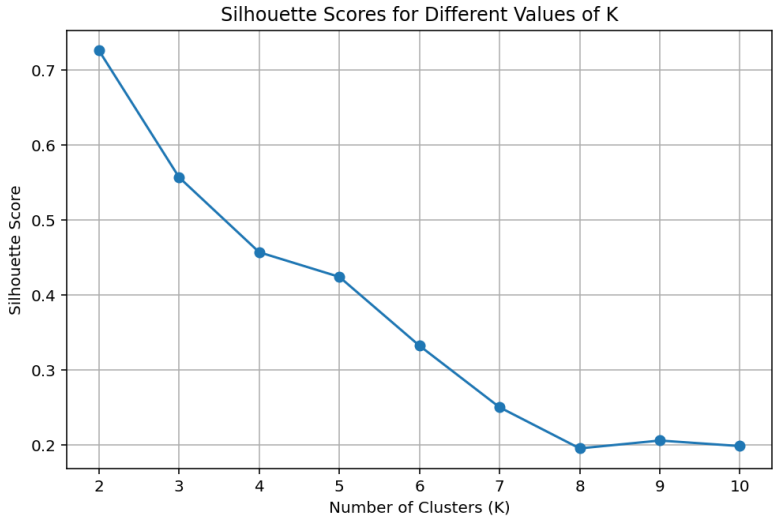
This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101119590

Clustering on Embeddings

- Clustering on Embeddings:
  - No clear elbow in the inertia curve → suggests gradual improvement without a sharp optimal K.
  - K-Means applied to GNN-generated embeddings showed a silhouette score of 0.73 for K=2 → strong separation between the two clusters.
  - Inertia of 2530, a 28% reduction from K=1 (inertia of 3501).
  - GNN has captured latent network structure that may not be visible in raw features.



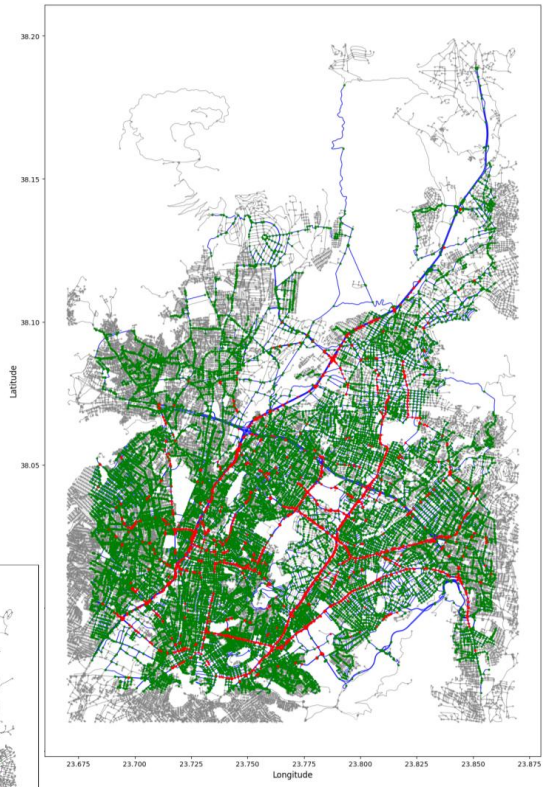


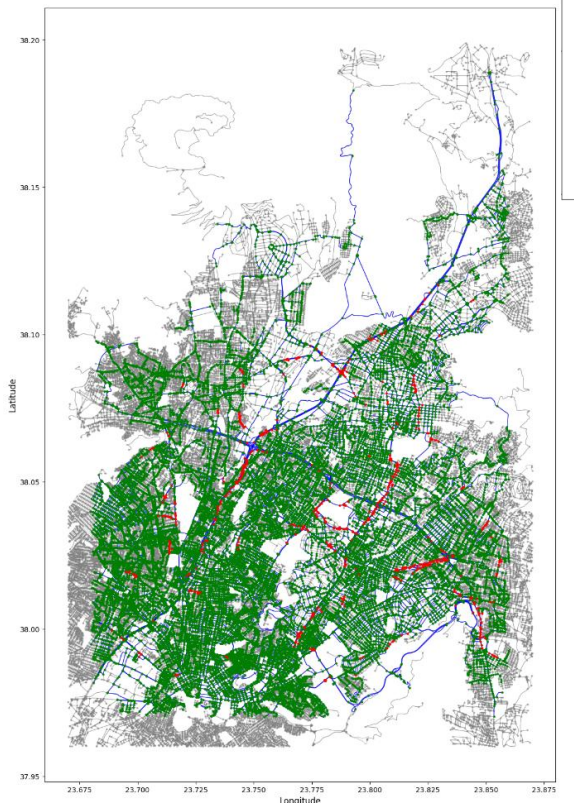


This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101119590

Clustering Comparison

- The groups identified by both clustering methods exhibit similar characteristics upon comparison.
- Green nodes represent safer behaviors, while red nodes are associated with less safe events and more trips.
- However, the clustering based on the embeddings tends to polarize the nodes, shrinking the less safe cluster while expanding the other one.





Clustering type	Safer nodes	Riskier nodes
Simple Clustering	24776	2732
Embeddings Clustering	26587	921

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

Assessing Clustering Results

➢ Each metric captures a different aspect of clustering quality:

- **WCSS**
  Measures intra-cluster compactness.
  ❖ *Lower values suggest tighter, more cohesive clusters.*

- **Silhouette**
  Evaluates how similar a point is to its own cluster vs others.
  ❖ *Higher values indicate well-separated, dense clusters.*

- **Davies-Bouldin Index (DBI)**
  Compares intra-cluster and inter-cluster distances.
  ❖ *Lower values reflect tighter and better separated clusters.*

- **Calinski-Harabasz Index (CHI)**
  Assesses cluster dispersion relative to the overall data spread.
  ❖ *Higher scores indicate well-separated, distinct clusters.*

Index	Simple Clustering value	Embeddings Clustering value
WCSS	145909	2530
Silhouette	0.58	0.73
DBI	1.41	0.98
CHI	8794	10551

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

$$CHI = \frac{Tr(B_k)}{Tr(W_k)}\frac{n-k}{n-1}$$

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

Overview of the Identified Groups

	Street_Count	Smoothened Speed	SpeedingFlag	Mobile_usage	Harsh_acc	Harsh_brk	Trips_count
Risky Cluster	3.15	47.93	19.78	6.42	0.38	0.33	92.2
Safer Cluster	3.34	28.55	0.92	2.27	0.09	0.08	20.24

- Mapped embedding cluster labels to raw data and averaged features per cluster to interpret differences.
  - Within the risky group of nodes, drivers tend to travel at higher speeds, with SpeedingFlag triggered on average 20 times more often.
  - There is also a higher prevalence of phone usage and harsh driving events.
  - These nodes are more heavily trafficked, as indicated by a greater number of trips.
  - In contrast, the safer group is characterized by lower traffic volumes and generally better driving behavior, including reduced speeds and fewer risky events.

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

Discussion

➢ Numerical and Conceptual Reflections

- The clear quantitative separation between groups validates the use of embedding-based clustering for identifying meaningful clusters. The magnitude of difference in key indicators (e.g., 20× speeding events) supports the practical relevance of the classification.
- The results align with known road safety principles: higher speeds and distracted driving increase risk.

➢ Methodology Suitability

- Clustering telematics and geometric features effectively identifies distinct areas within a road network.
- Reliance on aggregate averages may mask temporal variations.
- Integrating cluster labels with raw features and averaging values per cluster bridges the gap between complex representation learning and real-world feature insights, improving explainability.

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

Potential applications

- The work provides actionable insights, informing on where to focus safety efforts and resources, aiming to improve overall traffic management and public safety.
  - Risky cluster areas can be targeted for interventions, such as infrastructure improvements, awareness campaigns, or enforcement measures, to enhance road safety.
  - Insurers can use this clustering to define risk profiles by identifying patterns of risky or safe driving behavior.
  - Drivers in high-risk clusters may face higher premiums, while those in safer clusters could benefit from discounts.
  - This also enables insurers to offer more accurate, location-based pricing and targeted advice.

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

Conclusions

- Graph-based representations enhance understanding of complex road safety data.
- By incorporating node features, topology, and edge attributes clustering performance are improved.
- K-Means used as baseline for efficiency; alternative clustering methods could be explored.
- Future directions include testing different GNN architectures and loss functions.
- Adding traffic and temporal features may increase real-world applicability and impact.

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

Thank you for the attention

Simone Paradiso

IVORY PhD Candidate and Researcher

Together with:
George Yannis & Apostolos Ziakopoulos

Department of Transportation Planning and Engineering
National Technical University of Athens

oseven

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

IVORY