Exploring Safe and Eco Driving Behavior through Large-scale Data using Unsupervised Learning

Virginia Petraki

Transportation Engineer, Research Associate

Together with:
Dimitris Nikolaou and George Yannis

Department of Transportation Planning and Engineering
National Technical University of Athens

ICTR 2025

12th International Congress on Transportation Research

16-18 October 2025, Thessaloniki, Greece

HIT- HELLENIC INSTITUTE OF TRANSPORT

HELLENIC INSTITUTE OF TRANSPORTATION ENGINEERS

Introduction

- Road crashes remain a major and growing global challenge, contributing to approximately 1.19 million fatalities annually

- Simultaneously, the transport sector accounts for around 25% of the EU's total CO_2 emissions and 31% of its total energy consumption

- Driving behavior is one of the most critical factors affecting road safety and efficiency

- Research often classifies trips and drivers into profiles (e.g., aggressive, distracted, risky, eco-conscious, safe), but the relationship between eco-driving and safe-driving remains underexplored

Objectives

This study aims to explore the intersection of safe and eco-driving behavior using real-world trip data and unsupervised learning techniques

The following research questions guide the analysis:

Q1. How can trips be meaningfully clustered according to driving context?

Q2. What are the key behavioral parameters that characterize safe and eco-efficient driving styles?

Q3. Do safe and eco-driving behaviors consistently align?

Methodology

- Two-Level K-Means Clustering Approach

  Level 1: Trips segmented by route characteristics

  Level 2: Each cluster was further divided using behavioral indicators

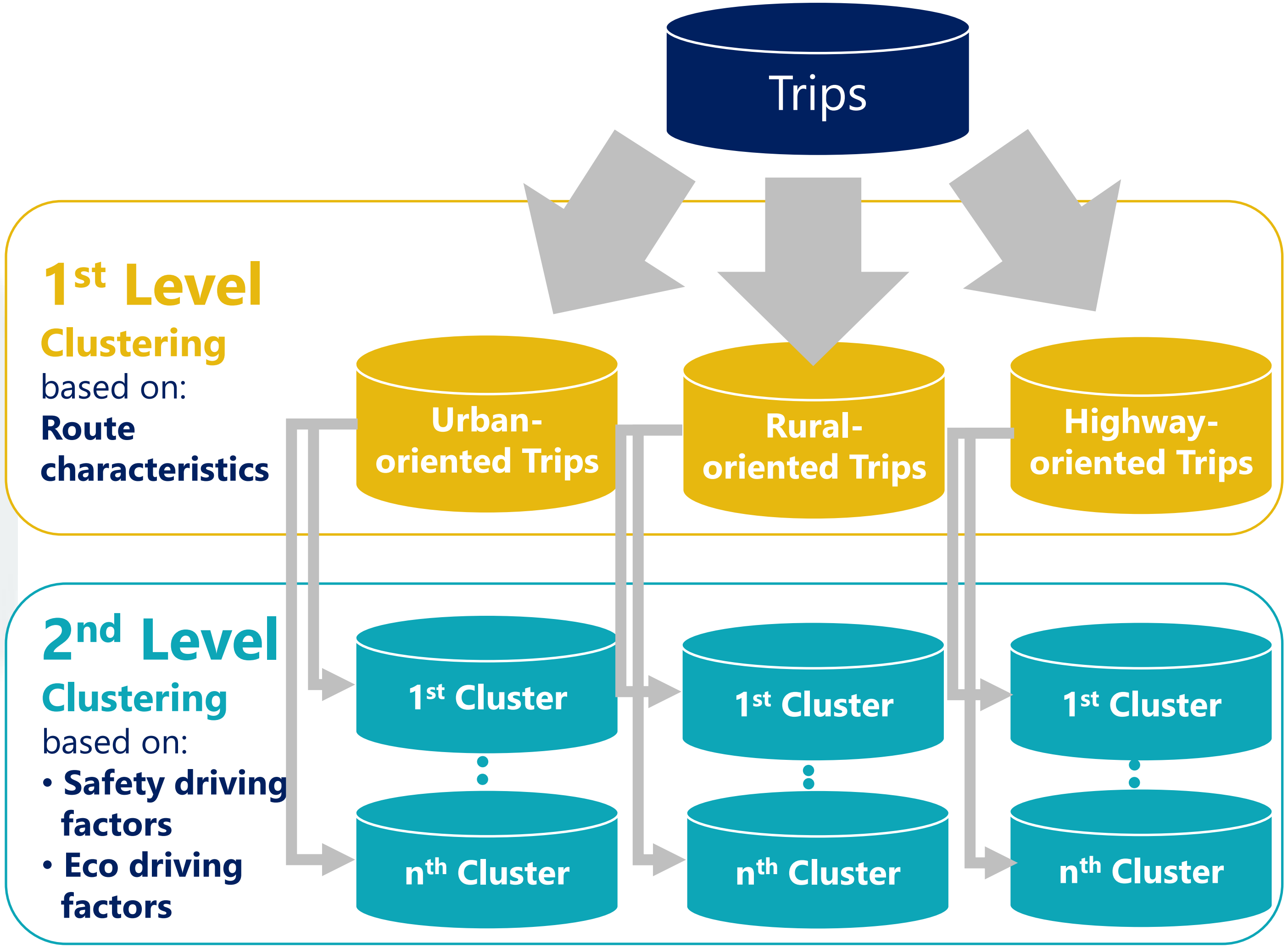
- Each clustering level followed a process:

(1) determining the optimal number of clusters using the Silhouette method,

(2) applying dimensionality reduction techniques (PCA), where appropriate,

(3) performing clustering using K-means,

(4) evaluating the classification quality.



Data Overview

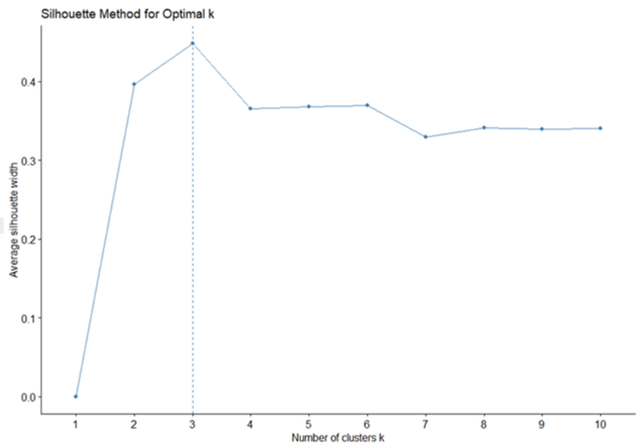
- The data was collected:
  - using smartphone sensors
  - provided by the OSeven Telematics Company
  - in an anonymized format
  - from 16,118 trips
  - over a 3-month period (March to May)
  - during the years 2023, and 2024
  - across Athens, Greece
- The 1st level clustering relies on variables such as the percentage of trip duration spent on urban, rural, and highway roads, and average trip speed

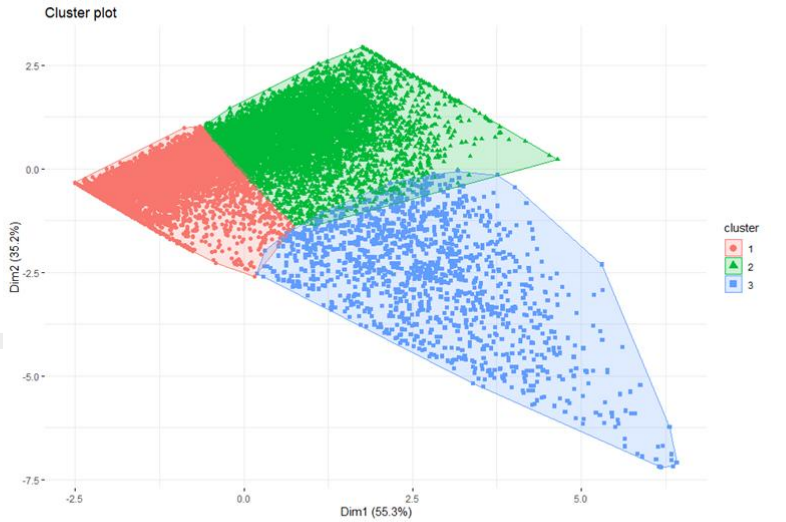
Variable Name	Description	Summary Statistics Min	Median	Q3	1st Level	2nd Level
Urban_prc	Percentage of trip duration spent on urban roads (Speed Limit<50)	0.00	59.56	81.98	●	
Rural_prc	Percentage of trip duration spent on rural roads (50<Speed Limit<80)	0.00	35.14	55.18	●	
Highway_prc	Percentage of trip duration spent on highways (Speed Limit>80)	0.00	4.16	0.00	●	
Speed_Avg	Average speed of the trip (km/h)	4.73	26.24	36.98	●	
Harsh_Brk_per_km	Number of harsh braking events per kilometer	0.00	0.08	0.27		●
MobilePhone_min_per100km	Minutes of mobile phone usage per 100 kilometers	0.00	0.00	6.410		●
Speed_Q90	90th percentile of speed during the trip	16.82	53.60	73.20		●
Acc_QCV	Coefficient of variation of acceleration (QCV= $100 \times \frac{Q_3-Q_1}{Q_3+Q_1}$)	0.00	0.60	0.64		●
Fuel_lit_per100km	Fuel consumption measured in liters per 100 kilometers	2.39	8.28	10.79		●

- The 2nd level clustering includes variables related to safety and fuel efficiency, like harsh braking events per kilometer, phone usage, high-speed tendencies (90th percentile of speed), variability in acceleration, and fuel consumption

1st Level Cluster Analysis

- 1st level clustering groups trips by road-type composition and average speed to represent the exogenous driving context (speed limits, geometry, traffic flow)
- Isolating this context upfront contributes to more interpretable Level-2 safety and eco clusters that reflect behavioral differences rather than environmental conditions
- 3 clusters with silhouette widths ranging from 0.35 to 0.52
- Cluster 1, is the largest and is characterized by a dominant share of urban road share and the lowest average speed
- Cluster 3, is the smallest and features the highest average speed and the highest highway share.



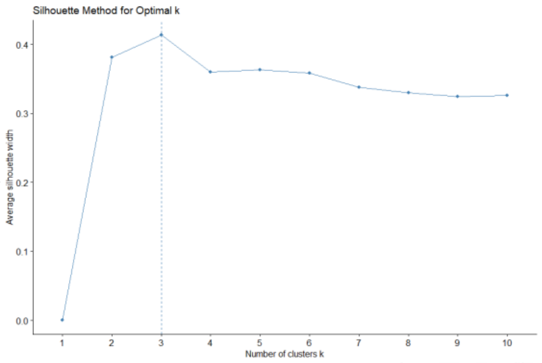


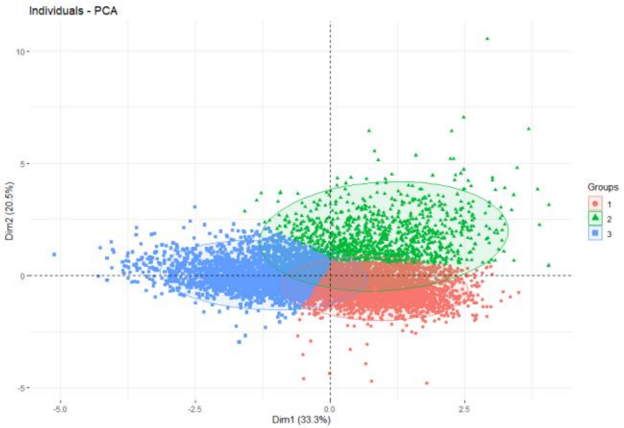
Cluster		Speed_Avg	Urban_prc	Rural_prc	Highway_prc	Trips	Silhouette width
1	Urban-Oriented	23.48	83.49	16.09	0.52	7,675	0.52
2	Rural-Oriented	31.23	39.32	59.25	1.33	7,145	0.39
3	Highway-Oriented	62.30	28.95	28.70	41.27	1,298	0.35

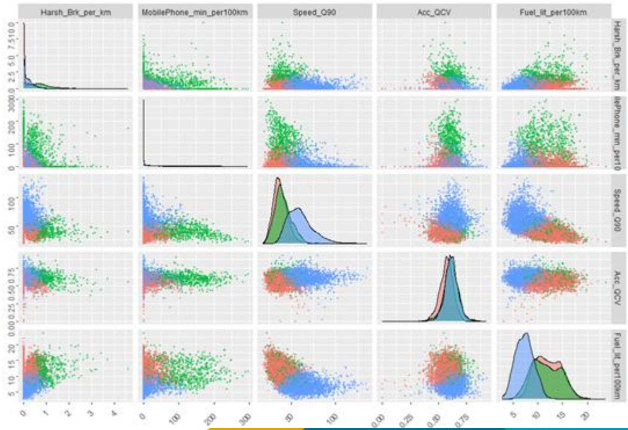
2^{nd} Level – Urban

- A 2^{nd} level clustering was performed for urban-oriented trips
- To enhance interpretability and reduce dimensionality, PCA was applied, retaining the first 2 principal components, which explained 54% of the variance
- Fuel consumption and speed factor are the most influential variables in PC1, while in PC2, harsh braking events and phone use are the dominant contributors
- Model choice: Silhouette curve favors k = 3
- Inefficient–Safe: Low surrogate crash risk (limited harsh braking and distraction) but high fuel intensity
- Inefficient–Risky: Elevated surrogate crash risk combined with inefficient fuel use
- Eco-Safe: Favorable safety and efficiency metrics despite higher Q90 speeds, potentially reflecting higher motorway or rural trip share or decreased congestion

Loadings	PC1 (33.4 %)	PC2 (20.5 %)
Harsh_Brk_per_km	0.09	**0.65**
MobilePhone_min_per100km	0.21	**0.58**
Speed_Q90	**-0.67**	0.09
Acc_QCV	-0.23	0.48
Fuel_lit_per100km	**0.67**	-0.02





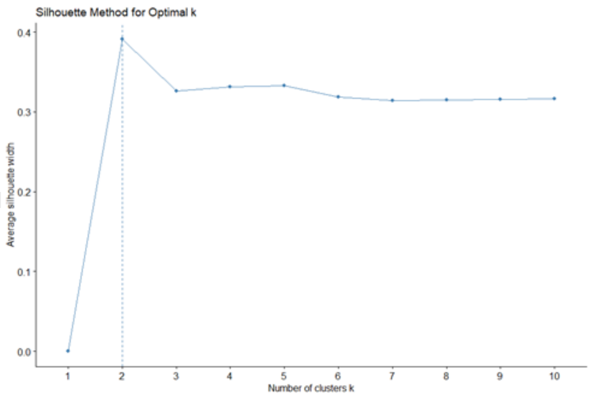


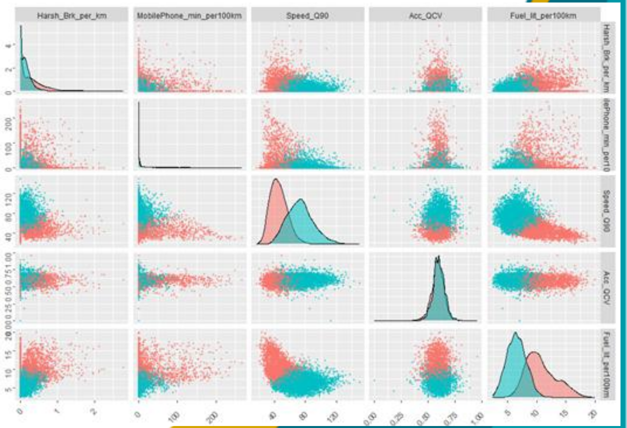
Cluster		Harsh_Brk_per_km	MobilePhone_min_per100km	Speed_Q90	Acc_QCV	Fuel_lit_per100km	Trips	Silhouette width
1	1_1 Inefficient-Safe	0.116	7.023	37.482	0.570	12.025	3,418	0.31
2	1_2 Inefficient-Risky	0.630	73.235	40.746	0.612	11.712	1,065	0.15
3	1_3 Eco-Safe	0.178	6.497	60.786	0.616	7.285	3,192	0.25

2nd Level - Rural

- A 2nd level clustering was performed for rural-oriented trips
- To enhance interpretability, PCA was applied; the first two PCs explain 55% of variance
- PC1 reflects an efficiency–speed axis (↑ fuel intensity, ↓ Q90 speed), while PC2 captures kinematic aggressiveness (↑ acceleration variability, ↑ harsh braking)
- Model choice: Silhouette curve favors k = 2
- Risky–Inefficient: Elevated surrogate crash risk (more harsh events and distraction), moderate Q90 speeds, and higher fuel use—characteristic of volatile, mixed-traffic rural segments
- Eco–Safe: Low distraction and harsh events with best fuel economy and higher Q90 speeds—consistent with steady, free-flow rural/motorway conditions

Loadings	PC1 (33.8 %)	PC2 (21.0 %)
Harsh_Brk_per_km	0.28	**0.60**
MobilePhone_min_per100km	0.33	-0.04
Speed_Q90	**-0.62**	0.11
Acc_QCV	-0.12	**0.79**
Fuel_lit_per100km	**0.64**	0.01





Cluster		Harsh_Brk_per_km	MobilePhone_min_per100km	Speed_Q90	Acc_QCV	Fuel_lit_per100km	Trips	Silhouette width
1	2_1 Risky-Inefficient	0.268	21.504	45.504	0.589	10.764	2,709	0.25
2	2_2 Eco-Safe	0.127	4.511	73.393	0.600	6.523	4,436	0.40

2nd Level - Highway

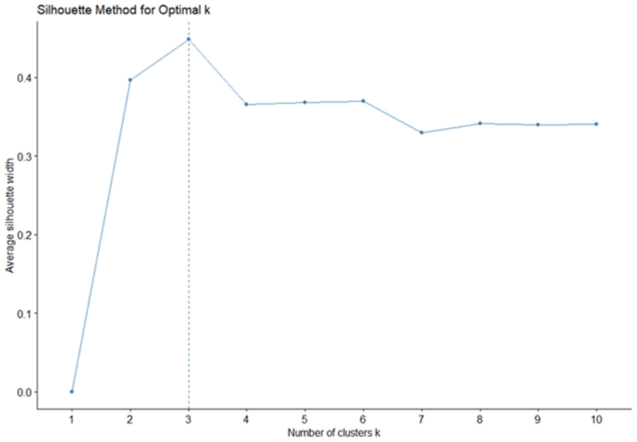
- A 2nd level clustering was performed for highway-oriented trips

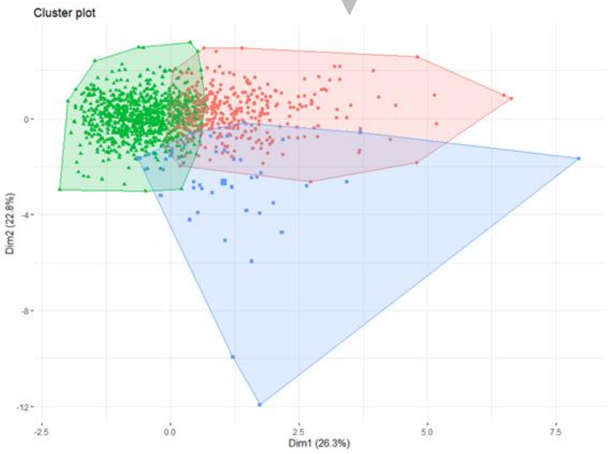
- Model choice: The silhouette curve favors k = 3

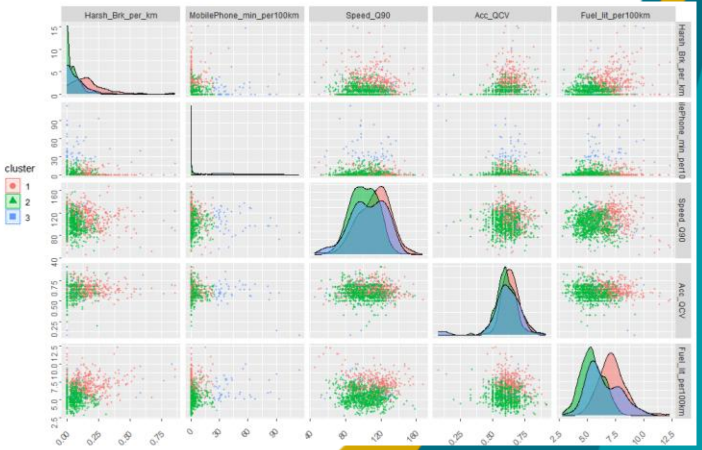
- Aggressive–Inefficient: Highest speeds, more harsh events and volatility, with the worst fuel economy

- Eco–Safe: Smooth, attentive cruising—lowest harsh braking, low phone use, and best fuel economy at high but steady speeds

- Distracted–Efficient: Very high phone use and moderate fuel use







Cluster		Harsh_Brk_ per_km	MobilePhone_ min_per100km	Speed _Q90	Acc _QCV	Fuel_lit _per100km	Trips	Silhouette width
1	3_1 Aggressive-Inefficient	0.170	1.959	111.866	0.645	7.229	375	0.20
2	3_2 Eco-Safe	0.043	1.649	100.989	0.605	5.442	876	0.30
3	3_3 Distracted-Efficient	0.089	44.881	105.287	0.602	6.433	47	0.22

Discussion

- The two-stage clustering framework identified **eight distinct driver-behavior profiles**
- **Eco–safety is context-dependent**. In steady-flow environments (rural/motorway), safety surrogates (low harsh events, low distraction) tend to align with fuel efficiency; in stop-and-go urban regimes, low-risk behavior can still be energy-inefficient due to congestion and idling
- Efficiency improves when moving from low→moderate **speeds** in urban/rural settings, but deteriorates at very high highway speeds

Conclusions

- This study leveraged a two-level K-means clustering framework to analyze over 16,000 real-world trips and identify **sustainable driving behaviors** across different road contexts
- The findings reveal that **the alignment between safe and eco-driving styles is not uniform** but varies significantly depending on the driving environment
- Shifting trips toward smooth, attentive, steady-state operation can deliver **dual benefits**—lower crash risk and reduced fuel/emissions—especially outside dense urban conditions
- **Cluster separation is moderate but actionable**. Silhouette values are mid-range, suggesting adequate for targeted interventions rather than hard classification
- **Future analysis** can quantify the safety–eco performance trade-off using supervised models or/and a segment-level analysis.

Exploring Safe and Eco Driving Behavior through Large-scale Data using Unsupervised Learning

Virginia Petraki
Transportation Engineer, Research Associate

Together with:
Dimitris Nikolaou and George Yannis

Department of Transportation Planning and Engineering
National Technical University of Athens

ICTR 2025

12th International Congress on Transportation Research

16-18 October 2025, Thessaloniki, Greece

HIT- HELLENIC INSTITUTE OF TRANSPORT

HELLENIC INSTITUTE OF TRANSPORTATION ENGINEERS