Eco-friendly Carpool Optimization Using Clustering and Travel Data

Amina Hromić, Nerma Kadrić,
Nadina Miralem, Azra Žunić
University of Sarajevo
Faculty of Electrical Engineering
Dept. of Computer Science and Informatics
Sarajevo, Bosnia and Herzegovina
ahromic1@etf.unsa.ba, nkadric1@etf.unsa.ba,
nmiralem1@etf.unsa.ba, azunic3@etf.unsa.ba

Anna Joliot, Philippe Canalda
Université de Franche-Comté
UFR STGI, Campus Marie et Louis Pasteur
Montbéliard, France
anna.joliot@femto-st.fr,
philippe.canalda@univ-fcomte.fr

***Abstract*—This paper presents a data-driven framework for optimizing eco-friendly carpooling using clustering algorithms and synthetic mobility data. The goal is to reduce CO_2 emissions and travel costs by intelligently forming carpool groups. A baseline greedy algorithm is compared with KMeans and AGNES clustering methods across multiple feature sets and cluster sizes ($k/n \in \{10, 30, 50, 60\}$), with finer granularity (e.g., $k = 50$) achieving optimal results. An optimization layer incorporates transport reassignment, substituting long car trips with trains when beneficial. Evaluation on a synthetic dataset of 300 missions (extended to 600/1,200 for scalability testing) across 7 French cities shows clustering-based solutions reduce CO_2 emissions by 27% (from 11,466 kg to 8,398 kg) compared to the greedy baseline, while maintaining stable costs. AGNES with $n = 60$ clusters yielded the highest Silhouette score (0.670), demonstrating superior cluster cohesion. The framework's scalability is validated on larger datasets, confirming its potential for sustainable institutional mobility systems.**

***Index Terms*—Carpooling, sustainable mobility, clustering, KMeans, AGNES, CO_2 reduction, optimization, Gower distance, scalability.**

I. Introduction

Transportation accounts for a significant share of global greenhouse gas emissions, with passenger cars alone responsible for nearly 60% of road transport-related CO_2 emissions in Europe [6]. As urban mobility grows and environmental concerns intensify, sustainable alternatives such as carpooling have emerged as promising mitigation strategies.

Carpooling not only reduces fuel consumption and traffic congestion, but also decreases the number of vehicles on the road [1]. However, the success of carpooling systems heavily depends on their ability to match users with compatible routes, schedules, and preferences—an inherently complex optimization problem.

Advances in data science have enabled more dynamic and personalized carpooling models, particularly through the use of unsupervised machine learning techniques such as clustering [3]. Unlike supervised learning, where models are trained on labeled data, unsupervised learning identifies patterns and structures in unlabeled datasets. Clustering is a key method within this paradigm, involving the automatic grouping of data points—in this case, travelers—based on feature similarity.

For carpooling, clustering enables the formation of traveler groups that share similar trip characteristics such as origin, destination, timing, cost, and environmental footprint. This facilitates scalable group formation without manual intervention. Gower distance is used to measure similarity across both numerical and categorical data, improving flexibility in modeling realistic travel behavior.

This paper proposes a carpool optimization framework that integrates clustering with post-processing logic for transport reassignment and cost minimization. A baseline greedy algorithm is compared against clustering-based methods (KMeans and AGNES), enhanced by an optimization layer that incorporates fallback strategies and substitutes long car trips with trains when appropriate.

Due to data privacy regulations such as the General Data Protection Regulation (GDPR), which governs personal data usage across the European Union, and limited access to institutional mobility records, a synthetic dataset was constructed using Python libraries such as `faker`, `random`, and `numpy`.

Evaluation focuses on emission reduction, group inclusion, cost-efficiency, and algorithm scalability. Results confirm that clustering methods, combined with domain-specific optimization, significantly outperform greedy baselines in terms of environmental and economic impact.

The remainder of the paper is organized as follows: Section II presents related work. Section III outlines dataset

generation and methodology. Section IV discusses experimental results. Section V concludes the work and provides directions for future research.

II. Related Work

Carpooling has long been recognized as a viable method to reduce transportation-related emissions and costs. Shaheen et al. [1] emphasize the ecological and economic benefits of shared mobility, citing reductions in energy usage, traffic congestion, and greenhouse gas emissions. Their work underscores the influence of technological trends—such as real-time ride-matching and route optimization—and demographic shifts that facilitate broader adoption of carpooling systems.

Complementing this, Das et al. [2] introduced a dynamic stock model to evaluate the long-term environmental impact of carpooling. Their model incorporates avoided trips and embedded emissions, offering a more holistic perspective. This motivated this paper's inclusion of vehicle characteristics and indirect emissions in our optimization calculations.

Pramanik [3] explored the application of both supervised and unsupervised machine learning techniques in carpooling optimization, addressing challenges such as route inefficiencies and low vehicle occupancy. While their work focuses on conceptual frameworks, this paper's study contributes a concrete implementation using clustering (KMeans and AGNES) with Gower distance and structured evaluation of optimization impact.

Chang et al. [4] presented a hybrid model for ride-hailing services that integrates XGBoost-based arrival prediction and reinforcement learning for vehicle dispatch. Although this project did not directly implement predictive time-series models or reinforcement learning due to computational limitations, in this paper, similar decision-making structure—prioritizing environmental gains are adopted through clustering and rule-based fallback strategies.

Campana et al. [5] focused on personalization through machine-learned ranking to enhance user satisfaction in carpooling services. While current framework in this paper is not user-personalized, their findings point toward a promising direction for incorporating user-centric variables such as preferences for departure times, co-passenger compatibility, or transport mode.

Collectively, these studies highlight the potential of intelligent systems to enhance the environmental efficiency of shared mobility. Work in this paper builds upon their foundations by integrating clustering with optimization logic, using a synthesized dataset modeled on real institutional travel data to evaluate the practicality and scalability of such approaches.

III. Methodology

The methodology combines synthetic data generation, clustering techniques, and a custom optimization layer to construct and evaluate carpooling scenarios under realistic academic travel constraints.

All experiments were implemented in Python 3.10 using `pandas`, `scikit-learn`, and `scipy` for clustering and data processing. Visualizations were produced using `matplotlib` and `seaborn`. The KMeans algorithm was configured with `n_init=10` and `random_state=42`, while AGNES used average linkage. All runs were performed on a standard workstation (Intel i7 CPU, 16 GB RAM), with typical experiment durations under 30 seconds. The full codebase and dataset generation scripts are available at GitHub.

A. Dataset Overview

The synthetic dataset was designed to emulate real-world academic travel patterns while addressing GDPR compliance constraints. Constructed using Python's `faker` and `numpy` libraries, it comprises three core components: (1) individual user profiles, (2) travel missions, and (3) vehicle emission factors.

The dataset initially contained 300 mission records across seven major French cities (Paris, Lyon, Strasbourg, Toulouse, Nice, Dijon, and Montbéliard), with subsequent validation on expanded sets of 600 and 1,200 missions to test scalability. Mission attributes include precise origin-destination pairs, datetime stamps (spanning May 2025), and transport modes (car, train, or plane). Approximately 40% of missions represent round trips, mirroring typical conference travel patterns.

Emission factors were derived from authoritative European Environment Agency (EEA) benchmarks, with CO_2 rates calculated per kilometer for each transport mode. Private vehicle emissions account for engine type variations (electric, hybrid, diesel), while train emissions reflect French regional (TER) and high-speed (TGV) averages. This granularity enables accurate environmental impact assessment during optimization.

Key dataset characteristics include:

- **Spatial Coverage**: 7 cities with real-world distances (Belfort → Paris: 410 km)
- **Temporal Distribution**: 11 travel days in May 2025 (peak on May 4th-6th)
- **Vehicle Mix**: 64% car ownership rate (capacity=5), with 32% electric/hybrid

The dataset generation process employed `pandas` for structuring and `datetime` for temporal logic, ensuring realistic scheduling constraints (e.g., ±2-day departure windows). This approach balances computational tractability with ecological validity, though future work

could incorporate real institutional travel logs when available.

B. System Architecture

The overall system architecture is shown in Figure 1. The pipeline begins with the integration of three primary datasets: individual missions, person profiles, and CO_2 emission factors. These are fed into the clustering module, where travelers are grouped using unsupervised algorithms (KMeans or AGNES). Clustering is based on a combination of trip features, such as distance, cost, date, and emissions. The resulting clusters are then passed to the optimization module, which forms valid carpool groups based on scheduling compatibility, car ownership, and vehicle capacity.

An additional layer evaluates transport reassignment logic, replacing car trips with trains if the train offers a substantially shorter travel time. The final output consists of optimized group assignments along with aggregated cost and emission metrics, enabling assessment of overall efficiency and sustainability.

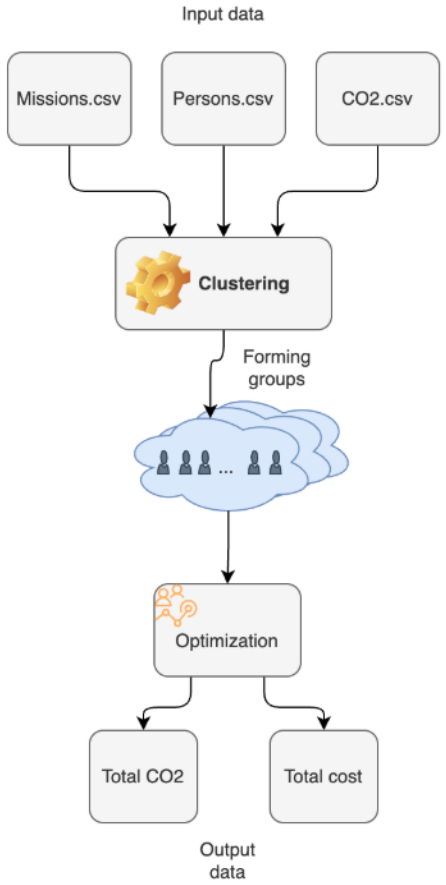

Figure 1: Global architecture of the carpool optimization system.

Figure 1 presents a high-level overview of the system pipeline. The diagram outlines how user data and emissions factors are preprocessed and fed into the clustering phase. It also shows how the optimization module builds upon clustering results to generate final carpool groups, incorporating substitution logic for trains and fallback mechanisms when necessary.

C. Baseline Greedy Algorithm

As a reference method, a greedy algorithm was implemented to form carpool groups by iteratively selecting a driver from the pool of users who own cars. Travelers with matching origin and destination within a two-day interval were assigned as passengers, up to the driver's capacity. Unassigned users defaulted to solo travel, using a fallback strategy based on vehicle availability. While simple and inclusive, this approach did not attempt to minimize emissions or costs globally.

D. Clustering-Based Grouping

To improve grouping efficiency, clustering was introduced as a pre-processing step. Travelers were embedded as feature vectors, using configurations ranging from four to eleven attributes, depending on the experiment. Clustering was performed using KMeans and AGNES, with Gower distance as the similarity metric to accommodate heterogeneous data. Dimensionality reduction via PCA enabled visual inspection of cluster compactness and separation. Different values for the number of clusters were explored, ranging from 10 to 60.

Although Gower distance is well-suited for datasets that contain both numerical and categorical features, it may introduce limitations in weighting interactions across diverse variable types. For instance, minor scale differences in numerical attributes could dominate similarity scoring if not normalized carefully. In future iterations, learned similarity functions or alternative composite metrics could further improve the precision of clustering in complex mobility datasets.

The selection of KMeans and AGNES was based on their complementary strengths. KMeans is efficient for larger datasets and provides straightforward cluster assignment, while AGNES enables hierarchical analysis and performs better with heterogeneous cluster shapes. Other algorithms, such as DBSCAN or HDBSCAN, were initially explored but exhibited suboptimal results due to variable density and high dimensionality in the synthetic dataset. Nonetheless, these methods remain promising for future extensions involving spatiotemporal or noise-prone data.

E. Optimization Layer

Following clustering, a rule-based optimization module was applied to each cluster to construct feasible carpool groups. Individuals were grouped based on route alignment, travel dates, and car availability. An extended

logic reassigned some car trips to train travel when rail offered a faster alternative by two or more hours. The final groupings minimized CO_2 emissions while ensuring that all users were assigned valid transport modes. Environmental impact was computed based on the specific emissions per kilometer of each mode, and travel costs incorporated accommodation and transport fees. A carbon tax was also applied to simulate the economic consequences of emissions under national policy.

F. Reinforcement Learning Attempt

During the initial design phase, reinforcement learning was explored as an alternative to static clustering. The carpool assignment problem was modeled as a Markov decision process, where states represented individual travelers and actions corresponded to potential group assignments. Q-learning was used to reward emissions reduction and group occupancy. However, the approach was excluded from the final implementation due to the extensive computation time and the complexity of reward tuning. Nonetheless, it remains a relevant direction for future real-time optimization frameworks.

IV. Evaluation and Results

The framework was evaluated through three clustering experiments comparing KMeans and AGNES algorithms, with the greedy algorithm as baseline. All metrics were calculated before and after optimization.

A. Baseline Greedy Algorithm

The greedy approach formed 197 groups from 300 missions, achieving full inclusion but with suboptimal emissions:

Table I: Greedy Algorithm Performance

Total CO_2 Emissions	11,466.47 kg
Total Travel Cost	€277,350
Groups Formed	197
Avg. Group Size	1.52

B. Clustering Experiments

Three experimental configurations were tested with cluster sizes $k, n \in \{10, 30, 50, 60\}$:

1) Experiment 1 (4 Features): Used distance (km), cost, CO_2/km, and return trip status.

Table II: Experiment 1 Results (Post-Optimization)

Algorithm	Clusters	CO_2 (kg)	Groups
KMeans	$k = 50$	8,563.72	291
AGNES	$n = 50$	8,447.94	293

Key Insight: AGNES achieved better emissions reduction (26.3% vs greedy) with comparable group counts.

2) Experiment 2 (11 Features): Expanded features including cities, vehicle types, and temporal components.

Table III: Experiment 2 Results (Post-Optimization)

Algorithm	Clusters	CO_2 (kg)	Groups
KMeans	$k = 50$	8,398.82	252
AGNES	$n = 50$	8,398.82	249

Key Insight: Both algorithms converged to identical emissions (26.7% reduction), suggesting feature saturation.

3) Experiment 3 (5 features): Used only cities, dates, and return trip status. Tested up to 60 clusters.

Table IV: Experiment 3 Results (Post-Optimization)

Algorithm	Clusters	CO_2 (kg)	Groups
KMeans	$k = 60$	9,758.29	161
AGNES	$n = 60$	9,806.33	148

Key Insight: Fewer groups formed but higher emissions due to omitted environmental features.

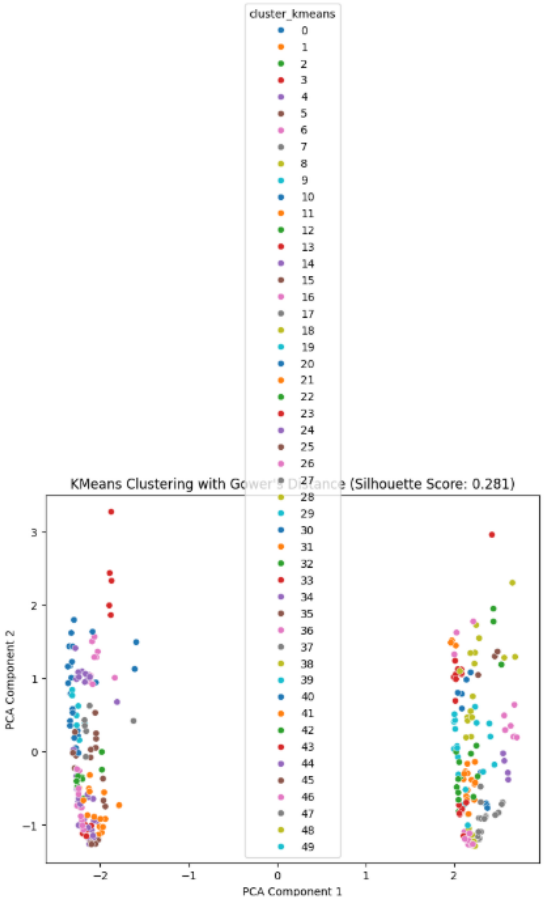

Figure 2: PCA visualization of KMeans clusters ($k = 50$) for Experiment 1, showing moderate separation.

C. Comparative Analysis

Table V demonstrates that clustering methods consistently outperform the greedy baseline, with KMeans (Exp2) achieving the highest emission reduction (26.7%) and AGNES (Exp1) showing superior cluster cohesion. Figure 3 visually confirms this advantage, with both algorithms maintaining stable improvements across multiple runs. The results validate that feature-rich clustering (Exp1-2) yields better environmental outcomes than spatiotemporal-only grouping (Exp3).

Table V: Optimal Configurations vs Baseline

Method	CO_2 (kg)	Reduction	Groups
Greedy	11,466.47	-	197
KMeans (Exp2)	8,398.82	26.7%	252
AGNES (Exp1)	8,447.94	26.3%	293

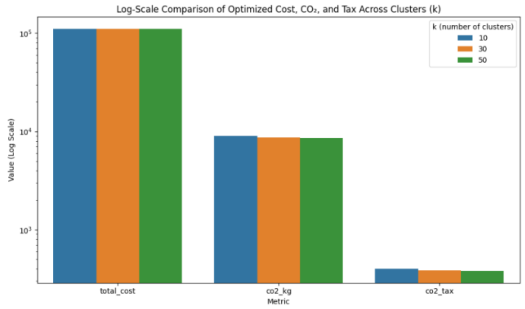

Figure 3: CO_2 emissions comparison for Experiment 1 configurations.

D. Effect of Train Substitution

One optimization strategy tested was the replacement of long-distance car trips with train travel where feasible. Train substitution was applied only when the train offered a travel time advantage of at least two hours. Figure 4 illustrates the impact of this substitution, showing a further decrease in overall emissions without compromising group inclusion.

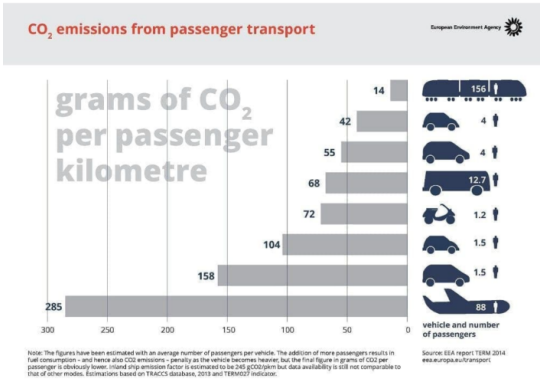

Figure 4: Impact of train substitution on total emissions.

E. Visualizing Final Carpool Groups

To illustrate the impact of clustering and optimization, Figure 5 presents the final carpool group formations for the best-performing configuration (AGNES, 50 clusters). This visualization confirms the algorithm's effectiveness in creating coherent and dense travel groups, balancing environmental and logistical constraints.

```
Cluster-based Carpool Evaluation — Source: /content/drive/MyDrive/Master IoT/Tutor project/Results/exp2_agnes50.csv
Total groups formed: 249
Total people assigned: 300
Total cost: €62020.0
Total CO₂ emission: 8398.82 kg
CO₂ tax: €374.59

Group group_1 | Cluster: 24
Route: Belfort → Strasbourg
Start Dates: 2025-05-03 — 2025-05-04
Travel Mode: car
Driver: Margot Berthelot (ID: 289)
(No passengers — solo driver or train traveler)

Group group_10 | Cluster: 3
Route: Montbéliard → Marseille
Start Dates: 2025-05-11 — 2025-05-11
Travel Mode: car
Driver: Stéphanie Ledoux (ID: 75)
(No passengers — solo driver or train traveler)

Group group_100 | Cluster: 9
Route: Belfort → Strasbourg
Start Dates: 2025-05-08 — 2025-05-08
Travel Mode: car
Driver: Laure Merle (ID: 33)
(No passengers — solo driver or train traveler)
```

Figure 5: Final carpool group formations after optimization using AGNES clustering with 50 clusters.

V. Conclusion

This work proposed a clustering-based framework for eco-friendly carpool optimization, evaluated on a synthetically generated dataset simulating academic travel in France. By combining unsupervised learning (KMeans and AGNES) with post-clustering optimization strategies, including transport reassignment logic, the framework achieved significant reductions in CO_2 emissions compared to a baseline greedy approach. The study highlighted the importance of feature selection, showing that even simple clustering configurations—when focused on relevant environmental attributes—can produce substantial impact. Among the evaluated techniques, AGNES offered slightly better performance in both clustering quality and environmental outcomes. Reinforcement learning was considered during early development stages but ultimately excluded due to computational constraints. Nonetheless, the proposed pipeline demonstrated efficiency, robustness, and potential scalability across larger datasets and extended time horizons.

VI. Future Work

Future directions include applying this framework to real-world datasets with live institutional travel information, enabling adaptive group formation in response to real-time constraints. Further enhancement could involve integrating user preferences for time, cost, or co-passenger matching, as well as expanding transport alternatives beyond cars and trains. Incorporating predictive components, such as time series forecasting or behavior-based recommendations, could increase the system's responsiveness to dynamic inputs. Additionally, deploying the pipeline as a mobile or web application would facilitate user interaction and operational

deployment. Finally, revisiting reinforcement learning under more favorable computational conditions could offer promising gains in adaptive and long-term optimization of shared mobility systems.

References

[1] S. Shaheen, A. Cohen, and A. Bayen, "The benefits of carpooling: Reducing emissions, congestion, and costs," in *Transportation Sustainability Research Center, UC Berkeley*, 2015.

[2] S. Das, P. Kalbar, and N. R. Velaga, "A dynamic stock model for estimating emissions from carpooling," in *Transportation Research Part D: Transport and Environment*, vol. 62, pp. 792–802, 2018.

[3] S. Pramanik, "Machine learning approaches for carpooling optimization," in *Journal of Transportation Technologies*, vol. 11, no. 1, pp. 1–15, 2021.

[4] W. Chang, H. Huang, and Y. Lee, "Emission-aware vehicle dispatching using XGBoost and reinforcement learning," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 5, pp. 3802–3813, 2022.

[5] P. Campana, F. Delmastro, and R. Bruno, "Personalized carpooling using machine-learned ranking," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 12, pp. 4561–4573, 2019.

[6] European Environment Agency, "Passenger transport emissions: CO emissions per km by mode," 2024. [Online]. Available: https://www.eea.europa.eu/

[7] ODYSSEE-MURE Project, "Energy efficiency indicators for transport in France," 2025. [Online]. Available: https://www.odyssee-mure.eu/

[8] GlobalPetrolPrices.com, "Gasoline prices in France," 2025. [Online]. Available: https://www.globalpetrolprices.com/France/gasoline_prices/