

Article

Data-Driven Clustering and Energy Characterization of Plug-In Hybrid Electric Vehicle Usage Patterns: A Gaussian Mixture-Based Framework

A. S. M. Bakibillah ^{1,*} , Md Abdus Samad Kamal ²  and Jun-ichi Imura ¹ 

¹ Department of Systems and Control Engineering, Institute of Science Tokyo, Tokyo 152-8552, Japan; imura@sc.e.titech.ac.jp

² Graduate School of Science and Technology, Gunma University, Kiryu 376-8515, Japan; maskamal@gunma-u.ac.jp

* Correspondence: bakibillah.a.aa@m.titech.ac.jp

Abstract

While plug-in hybrid electric vehicles (PHEVs) can significantly reduce fuel consumption and emissions, their real-world benefits strongly depend on heterogeneous driver usage patterns. Understanding these usage patterns is therefore essential for optimizing energy management and electrification policies. This study presents a data-driven framework for identifying and characterizing PHEV driving behavior using two primary indicators: the Utility Factor (UF), which quantifies the proportion of electric-mode driving, and the annual vehicle kilometers traveled (VKT), which measures driving intensity. A Gaussian Mixture Model (GMM) is employed in a transformed feature space characterized by the logit of UF and the logarithm of VKT to capture nonlinear relationships and diverse usage patterns. The Bayesian Information Criterion (BIC) is used to find the optimal number of behavioral clusters. To assess the robustness of the clustering, we perform a bootstrap stability analysis and compare it with k-means and a density-based clustering method (DBSCAN). Based on an analysis of real-world PHEVs, three distinct usage patterns are identified. The dominant cluster (96.6%) exhibits moderate electric usage (UF = 0.38) with an annual mileage of 18,617 km and fuel consumption of 3.89 L/100 km, whilst two smaller clusters represent near-full electric operation (2.3%, UF about 1.0, negligible fuel) and low-mileage users (1.1%, approximately 2444 km/year). The clustering demonstrates high assignment confidence (posterior entropy $< 10^{-4}$) and moderate stability, with a mean Adjusted Rand Index (ARI) of 0.448. The findings indicate that the proposed probabilistic clustering framework offers a comprehensible and statistically robust method for identifying diverse PHEV usage patterns. These insights can support adaptive energy management strategies, effective planning of charging infrastructure, and evidence-based policies to maximize the real-world electrification benefits of PHEVs.



Academic Editor: Emilio Larrodé

Received: 27 April 2026

Revised: 24 June 2026

Accepted: 25 June 2026

Published: 7 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: plug-in hybrid electric vehicles; Gaussian mixture model; vehicle kilometers traveled; utility factor; sustainable mobility

1. Introduction

As global transportation demand increases, the sector remains a major contributor to energy consumption and greenhouse gas emissions [1]. In the U.S., road transportation alone accounts for 28% of national energy use [2] and 29% of GHG emissions [3]. To meet increasingly stringent climate targets and reduce dependence on fossil fuels, vehicle

electrification has emerged as a critical pathway toward sustainable mobility [4–6], and many countries are introducing policies to support the development of electromobility [7,8]. Specifically, Plug-in Hybrid Electric Vehicles (PHEVs) serve as a pragmatic bridge technology that combines extended range capability with zero-emission potential [9,10]. PHEVs have experienced substantial growth in global markets over the past decade, though their adoption trajectory varies considerably across regions and has shown signs of stabilization in recent years. Global PHEV sales reached 14.2 million units in 2023, representing a 35% increase from the previous year, with PHEVs and battery electric vehicles (BEVs) combined accounting for 15.8% of all light-duty vehicle sales globally [11]. However, regional market dynamics reveal significant disparities: in the first half of 2022, PHEVs constituted 9% of new passenger car sales in Europe, 5% in China, and only 1.4% in the United States [12]. Despite this growth, PHEVs face criticism for real-world fuel consumption and carbon dioxide emissions that are two to four times higher than rated values when owners rely predominantly on combustion power rather than frequent charging [13]. Moreover, their GHG emissions reduction potential depends strongly on actual usage patterns and the carbon intensity of the underlying electricity generation.

A critical challenge in assessing the energy-saving potential of PHEVs lies in the substantial discrepancy between laboratory-certified efficiency ratings and real-world performance [14]. Jeong et al. [15] analyzed PHEV energy consumption and emissions across varied real-world driving conditions and highlighted that real-world operation can significantly alter expected environmental benefits compared with laboratory labels. Standardized test cycles, conducted under controlled conditions, systematically fail to capture the variability inherent in actual driving environments. Real-world measurements consistently demonstrate that factors such as ambient temperature, traffic congestion, driving style, and auxiliary loads significantly influence both electric range and fuel consumption [16,17]. For instance, Zhou et al. conducted on-road testing of two PHEVs in Beijing under diverse operating conditions, finding that the actual all-electric range could be reduced by at least 35% relative to manufacturer claims, with the worst-case scenario yielding only 14.1 km compared to the claimed 20 km [18]. Their study revealed that air-conditioning use, congested traffic, and higher loading substantially increased energy consumption and shortened the charge-depleting range, whilst real-world fuel consumption in charge-sustaining mode varied considerably from 3.5 to 6.3 L/100 km, depending on usage conditions. The substantial variability in real-world PHEV performance is well-documented across multiple geographic regions and vehicle models. Frey et al. [19] demonstrated through extended measurements that trip characteristics, ambient conditions, and driver behavior significantly influence operational energy use and emissions, whilst Graver et al. [20] confirmed that in-use performance consistently deviates from certification values. Mandev et al. [21] further identified charging frequency, trip distance distribution, and ambient temperature as critical determinants of actual fuel consumption in European PHEV fleets. Graba et al. [22] found wide variation in total energy consumption under actual operating conditions (reported range 20.19 to 41.97 kWh/100 km across electric and ICE operating modes) and emphasized dependence on usage and available energy in the fuel tank and traction battery.

Beyond performance variability, heterogeneity in PHEV usage patterns among users influences energy-saving potential. Williams et al. [23] demonstrated that the energy consumption and greenhouse gas emissions of PHEVs varied substantially depending on whether the vehicle was used as a primary or secondary household vehicle and on individual users' specific travel patterns. Recognizing this interdependence, subsequent research has advanced toward integrated optimization frameworks. Xu et al. [24] developed an integrated optimization framework that simultaneously addresses battery sizing, charging

scheduling, and power management in PHEVs, recognizing that these three aspects are intrinsically coupled. Xie et al. [25] evaluated coordinated management of connected plug-in hybrid electric buses for energy saving, inter-vehicle safety, and battery health. Li et al. [26] developed a bi-level predictive energy management strategy that uses convex optimization to simultaneously minimize fuel consumption and battery degradation.

Numerous works have proposed AI-driven methods for transportation electrification and clustering. Sun et al. [27] proposed a hybrid energy management strategy that combines Deep Learning for high-level power demand prediction with an improved Model Predictive Control (MPC) framework to optimize real-time power distribution and enhance fuel economy in PHEVs. Han et al. [28] proposed a deep reinforcement learning (DRL)-based energy management strategy (EMS) for plug-in hybrid heavy-duty trucks that combines driving-cycle pattern recognition with deep transfer learning and reported faster convergence (nearly 50%) and substantial reductions in battery degradation rate and overall driving cost. Zhang et al. [29] developed an adaptive weight-based predictive EMS that predicts optimal State of Charge (SOC) using cascaded neural networks, generates Pareto fronts via improved differential evolution, and embeds adaptive weight adjustment into an MPC framework to balance battery life and fuel economy. Dang et al. [30] proposed an energy management optimization framework for PHEVs that integrates the microgrid power supply, vehicle physical constraints, and real-time traffic flow to minimize total energy costs and carbon emissions through a multi-dimensional coupling approach. Recently, Doikin et al. [31] developed a machine learning (ML)-based control strategy that learns predictive vehicle usage behavior from real-world test data to inform PHEV control, reporting average fuel consumption between 3 and 4.2 L/100 km. Moreover, unsupervised and self-supervised ML were proposed to discover driving events, cluster driver types, and support charging-session simulation and range prediction [32,33].

Realizing the benefits of PHEVs depends critically on consumer usage behavior and policies that influence it, as well as on vehicle efficiency under diverse operating conditions. While PHEVs have attracted substantial research attention, understanding their real-world usage behavior remains limited. Existing studies primarily evaluate PHEV performance using aggregate energy consumption or emission metrics, often overlooking driver heterogeneity and variability in electric-mode utilization. In particular, the utility factor (UF) and annual vehicle kilometers traveled (VKT) are typically analyzed in isolation, limiting the ability to capture integrated behavioral patterns of PHEV usage. Moreover, conventional descriptive or regression-based approaches may not adequately capture latent behavioral regimes within heterogeneous vehicle populations. Although clustering techniques have been applied in transportation research, their use for PHEV usage characterization remains scarce, particularly frameworks that combine probabilistic clustering with appropriate feature transformations to account for bounded variables and nonlinear relationships. Additionally, few studies systematically assess clustering robustness and interpretability through multiple validation metrics.

This study addresses the above gaps by proposing a statistically rigorous, data-driven approach to identify latent PHEV usage patterns and correlate electric driving behavior, travel intensity, and fuel consumption. Specifically, we develop a Gaussian Mixture Model (GMM) in a transformed feature space comprising the logit of UF and the logarithm of VKT to capture nonlinear relationships and identify heterogeneous PHEV usage patterns. We determine the optimal number of clusters using the Bayesian Information Criterion (BIC), evaluate clustering robustness via bootstrap stability analysis, and compare with k-means and DBSCAN. The proposed probabilistic clustering framework delivers an interpretable and statistically rigorous identification of heterogeneous PHEV usage regimes. The main contributions are summarized as follows:

- Development of a data-driven behavioral clustering framework: We propose a probabilistic framework for identifying latent PHEV usage regimes using two crucial behavioral indicators, UF and VKT, thereby enabling systematic characterization of heterogeneous real-world driving patterns. Unlike prior studies, this work links behavioral clustering with energy consumption outcomes and policy-relevant electrification potential.
- Feature transformation for robust clustering of behavioral variables: We introduce a transformed feature space using the logit transformation of UF and the logarithm of VKT that allows bounded and highly skewed variables to be modeled effectively within the proposed GMM framework.
- Identification of distinct PHEV operational regimes: Our proposed approach reveals three behavioral clusters representing moderate electric usage, highly electrified driving, and low-mileage usage patterns, providing quantitative insights into the diversity of PHEV operational behavior. By linking UF, VKT, and fuel consumption across clusters, we provide an interpretable characterization of how different driving patterns influence real-world fuel usage and electric driving shares.

Beyond individual-level behavioral characterization, this study contributes to a system-level understanding of electrified transportation by linking heterogeneous driver usage patterns to aggregate energy demand and infrastructure utilization. By identifying latent behavioral regimes, the proposed framework enables the translation of micro-level driving behavior into macroscopic insights relevant to transportation system planning, charging infrastructure deployment, and energy system integration. Such a connection is essential for designing scalable strategies that enhance electric driving utilization and improve the overall efficiency of PHEV-dominated transport systems.

The paper is organized as follows. Section 2 describes the methodology, including data description, feature transformation, model formulation with parameter estimation, model selection, and post-clustering characterization. Section 3 explains the model validation and evaluation metrics. Section 4 presents key results with policy implications. Finally, Section 5 gives the conclusions and future research directions.

2. Methodology

This study analyzes data from 175 privately owned plug-in hybrid electric vehicles (PHEVs) in the German car fleet [34]. The original dataset contained a larger number of records; however, only observations with complete information for the variables required in the clustering and post-clustering analyses were retained. After removing records with missing values, 175 complete vehicle-level observations remained. Therefore, the dataset should be interpreted as an observational real-world sample rather than a statistically representative national fleet sample.

The sample comprises four models: Toyota Prius PHEV (89 vehicles), Mitsubishi Outlander PHEV (46 vehicles), Opel Ampera (25 vehicles), and Volvo V60 PHEV (15 vehicles). Let each observation $i \in \{1, 2, \dots, N\}$ represent the usage profile of a single plug-in hybrid electric vehicle (PHEV). Four quantitative variables are recorded, such as the utility factor $UF_i \in [0, 1]$, which measures the proportion of total driving distance conducted in electric mode; the annual vehicle-kilometers traveled $VKT_i > 0$, denotes the total driving distance accumulated within one year; the observation period $Days_i > 0$, is the duration of data collection in days; and the fuel consumption $F_i > 0$, given in liters per 100 km. The clustering analysis uses UF and VKT because they represent two central behavioral dimensions of PHEV usage: electric-mode utilization and annual travel intensity. UF measures the share of driving performed electrically, while VKT captures overall mobility demand. The variables $Days_i$ and F_i are used in post-clustering analyses of temporal coverage and energy

efficiency. Other factors, such as charging frequency, ambient temperature, vehicle model, and battery capacity, are not consistently available for all vehicles and therefore excluded to avoid missing-data bias. Table 1 summarizes the descriptive statistics of the complete vehicle-level dataset used in this study.

We applied various preprocessing steps before modeling to ensure statistical robustness. Specifically, the UF values are constrained to a narrow open interval $[\epsilon, 1 - \epsilon]$ with $\epsilon = 10^{-6}$ to avoid singularities (infinite values) when applying the logit transformation. All records with negative VKT_i or $Days_i$ are removed to ensure physical consistency. Additionally, to mitigate the influence of high-mileage outliers, the top and bottom 1% of VKT_i values are optionally winsorized. The resulting dataset $\mathcal{D} = \{UF_i, VKT_i, Days_i, F_i\}$ provides a clean, physically meaningful dataset suitable for subsequent analysis and Gaussian mixture-based clustering.

Table 1. Descriptive statistics of the PHEV dataset used in this study.

Variable	N	Mean	Std. Dev.	Median	Min	Max
Annual VKT (km)	175	20,451.63	10,773.11	19,107.00	1903.00	63,906.00
Utility factor (UF)	175	0.423	0.239	0.409	0.005	1.000
Observation days	175	504.63	1155.63	357.00	18.00	14,865.00
Fuel consumption (L/100 km)	175	3.83	1.65	3.85	0.00	8.06

In Figure 1, the overall workflow of the proposed GMM-based framework for PHEV usage characterization is presented. The raw PHEV data are first preprocessed and transformed into a feature space defined by the logit of UF and the logarithm of VKT. A full-covariance GMM is then estimated using the EM algorithm, and the optimal number of clusters is selected using BIC. The resulting clusters are evaluated using posterior probabilities, entropy, Mahalanobis distance-based geometry, and standard validation metrics, and then interpreted in terms of behavioral patterns, fuel consumption, and policy relevance.

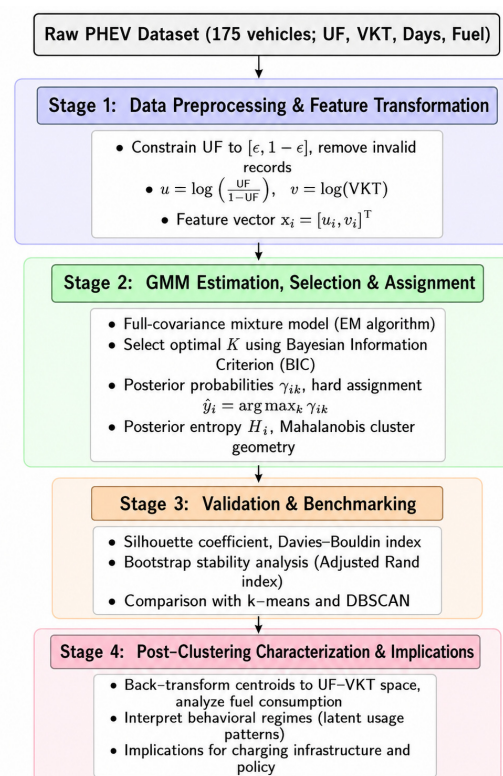


Figure 1. Overview of the proposed GMM-based framework for PHEV usage characterization.

2.1. Feature Transformations

In this study, the key behavioral indicators, UF and VKT, exhibit distinct statistical properties that require appropriate transformations before clustering. We applied monotone and variance-stabilizing transformations to map these variables into a continuous Euclidean space suitable for Gaussian-based modeling. Specifically, the UF variable is converted through a logit transformation, and VKT through a natural logarithmic transformation, expressed as

$$u_i = \log\left(\frac{UF_i}{1 - UF_i}\right), \quad v_i = \log(VKT_i), \quad (1)$$

where u_i and v_i denote the transformed UF and VKT for observation i . These transformations serve two purposes, i.e., (i) they linearize nonlinear relationships, and (ii) stabilize variances across the range of observed values. Then, each vehicle observation is represented as a two-dimensional feature vector as

$$x_i = \begin{bmatrix} u_i \\ v_i \end{bmatrix} \in \mathbb{R}^2, \quad (2)$$

which forms the basis for probabilistic clustering in the transformed space. The inverse mappings are used to interpret and visualize the cluster centroids on the original behavioral scales, given as

$$UF_i = \frac{1}{1 + e^{-u_i}}, \quad VKT_i = e^{v_i}. \quad (3)$$

This bidirectional transformation retains interpretability and ensures statistical regularity, allowing the Gaussian mixture model to operate effectively on well-conditioned data distributions. Note that transformations are not intended to establish exact univariate normality of the data. Instead, they improve the suitability of the feature space for Gaussian mixture modeling by mapping the bounded UF variable to the real line and reducing the skewness and scale imbalance of VKT. Since the GMM represents the data as a mixture of local Gaussian components, global normality of the full dataset is not required.

2.2. Model Formulation and Parameter Estimation

Based on the transformed feature space, we develop a probabilistic Gaussian mixture model (GMM) to detect latent driving-behavior regimes within the population [35]. The GMM is selected because it provides a probabilistic model-based clustering framework that can represent heterogeneous behavioral regimes with different covariance structures. Unlike k-means, which assigns each observation to the nearest centroid and assumes approximately spherical clusters, a full-covariance GMM can capture anisotropic and correlated structures in the transformed UF-VKT space. We assume that the empirical distribution of x_i originates from a finite mixture of K Gaussian components, each representing a distinct usage pattern characterized by a specific balance between electric-mode operation and overall driving intensity. The probability density function of the model is expressed as

$$p(x_i | \Theta) = \sum_{k=1}^K \pi_k \mathcal{N}(x_i | \mu_k, \Sigma_k), \quad \sum_{k=1}^K \pi_k = 1, \quad \pi_k > 0, \quad (4)$$

where π_k denotes the mixing proportion (prior probability) of cluster k , $\mu_k \in \mathbb{R}^2$ is the mean vector describing the center of cluster k , and Σ_k is the positive-definite covariance matrix that captures the spread and correlation between u_i and v_i within that regime. The full parameter set is given as $\Theta = \{\pi_k, \mu_k, \Sigma_k\}$.

To express the unobserved cluster memberships, we employ a latent binary variable $z_{ik} \in \{0, 1\}$, where 1 represents whether observation i belongs to component k and 0 otherwise. In this latent-variable basis, the complete-data log-likelihood can be expressed as

$$\mathcal{L}(\Theta) = \sum_{i=1}^N \sum_{k=1}^K z_{ik} [\log \pi_k - \frac{1}{2} \log |2\pi \Sigma_k| - \frac{1}{2} (x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k)]. \quad (5)$$

Because the membership indicators z_{ik} are unknown, parameter estimation is performed using the Expectation–Maximization (EM) algorithm [36,37], which iteratively maximizes the log-likelihood of the observed data.

In the E-step, the posterior probability γ_{ik} that observation i is associated with cluster k is computed using the current parameter estimates $\Theta(t)$ as

$$\gamma_{ik} = \Pr(z_{ik} = 1 \mid x_i, \Theta(t)) = \frac{\pi_k(t) \mathcal{N}(x_i \mid \mu_k(t), \Sigma_k(t))}{\sum_{j=1}^K \pi_j(t) \mathcal{N}(x_i \mid \mu_j(t), \Sigma_j(t))}. \quad (6)$$

Here, the numerator corresponds to the joint probability that observation i comes from component k , and the denominator is the total probability of observation i under the entire mixture model, obtained by summing over every possible component j (from 1 to K).

In the M-step, the posterior weights are utilized to update the model parameters by maximizing the expected log-likelihood as

$$N_k = \sum_{i=1}^N \gamma_{ik}, \quad \pi_k(t+1) = \frac{N_k}{N}, \quad \mu_k(t+1) = \frac{1}{N_k} \sum_{i=1}^N \gamma_{ik} x_i, \quad (7)$$

$$\Sigma_k(t+1) = \frac{1}{N_k} \sum_{i=1}^N \gamma_{ik} (x_i - \mu_k(t+1))(x_i - \mu_k(t+1))^T. \quad (8)$$

The above steps are alternated until convergence, which occurs when the relative change in log-likelihood falls below a threshold, e.g., 10^{-6} . Note that using a full covariance matrix Σ_k allows each cluster to adopt a flexible elliptical shape [38], thereby capturing potential correlations between UF and VKT within behavioral groups. This GMM formulation provides a probabilistic and interpretable framework for identifying latent driving-behavior regimes in the transformed (u, v) space.

The computational cost of the EM-based GMM estimation scales approximately as $O(I_{EM} N K d^2)$, where I_{EM} is the number of EM iterations and d is the feature dimension. In this study, $d = 2$ and the selected number of components is small, making the method computationally efficient. For larger transportation datasets, scalability can be improved using mini-batch EM, simplified covariance structures, parallel computation, or distributed implementation.

2.3. Model Selection and Cluster Assignment

After estimating the Gaussian mixture parameters, we determine the optimal number of latent clusters K that best represent the heterogeneity in PHEV usage patterns. Since the Gaussian mixture likelihood increases monotonically with K , a penalized information criterion is required to balance model fit and complexity. Therefore, the Bayesian Information Criterion (BIC) [39,40] is employed for model-order selection because it penalizes unnecessary model complexity and reduces the risk of overfitting when increasing the number of mixture components. The BIC is given by

$$\text{BIC} = -2\hat{\mathcal{L}} + p \log N, \quad (9)$$

where $\hat{\mathcal{L}}$ denotes the maximized log-likelihood of the fitted mixture model and p is the number of free parameters. For a two-dimensional Gaussian mixture with full covariance matrices, the number of parameters is given by

$$p = (K - 1) + 2K + 3K, \quad (10)$$

where the first term is the mixing weights (subject to the unit-sum constraint), the second term denotes two mean parameters for each component, and the third term corresponds to three distinct elements of each covariance matrix. The model with the lowest BIC is chosen as the optimal cluster number $\hat{K}$, which achieves the best trade-off between goodness of fit and model complexity. Then, each observation i is assigned to a cluster based on the posterior probability γ_{ik} obtained from (6). The cluster level for observation i is defined as

$$\hat{y}_i = \arg \max_{k \in \{1, \dots, \hat{K}\}} \gamma_{ik}. \quad (11)$$

The uncertainty associated with each observation i is quantified by its posterior entropy H_i as

$$H_i = - \sum_{k=1}^{\hat{K}} \gamma_{ik} \log(\gamma_{ik}), \quad H_i \in [0, \log \hat{K}], \quad (12)$$

where higher entropy values signify boundary points between behavioral regimes, while lower values correspond to confidently assigned observations. The geometric structure of each cluster in the transformed (u, v) space is further analyzed by the Mahalanobis distance [41]. Specifically, for observation i belonging to cluster k , the squared Mahalanobis distance d_{ik}^2 is given by

$$d_{ik}^2 = (x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k), \quad (13)$$

which follows a chi-squared distribution $d_{ik}^2 \sim \chi_2^2$ with two degrees of freedom under the Gaussian assumption. Contours associated with constant $d_{ik}^2 = c$ indicate ellipses that encompass regions of equal probability density in the transformed (u, v) space. These ellipses, when projected onto the original UF–VKT scale, provide a geometric visualization of the behavioral regimes identified by the model. Note that the above Gaussian-mixture analysis describes how clusters are determined; however, the interpretation of clusters and linking to meaningful physical quantities is necessary, as described in the next section.

2.4. Post-Clustering Characterization

Following the clustering procedure, each vehicle observation i is probabilistically assigned to one of the $\hat{K}$ latent components, thus identifying groups of drivers with similar behavioral profiles in the transformed (u, v) space. To interpret these clusters on the original physical scale, the component means $\mu_k = [\mu_{k,1}, \mu_{k,2}]^T$ are back-transformed using the inverse mappings introduced in (3) as

$$\overline{\text{UF}}_k = \frac{1}{1 + e^{-\mu_{k,1}}}, \quad \overline{\text{VKT}}_k = e^{\mu_{k,2}}, \quad (14)$$

where the subscript $k, 1$ denotes the coefficient for variable 1 (UF) in cluster k and $k, 2$ denotes the coefficient for variable 2 (VKT) in that same cluster. These quantities represent the characteristic electric driving share and annual distance associated with behavioral regime k . The within-cluster variability in UF and VKT is obtained from the associated covariance matrix Σ_k . Furthermore, variables not directly involved in the clustering process, namely the observation period Days_i and fuel consumption F_i , are examined post hoc to determine the operational and energetic characteristics of each cluster.

To investigate how energy efficiency varies across behavioral regimes, a simple cluster-specific regression is performed on the logarithm of fuel consumption and the transformed behavioral features. For cluster k , this relationship can be expressed as

$$\log F_i = \alpha_k + \beta_{k,1}u_i + \beta_{k,2}v_i + \varepsilon_{ik}, \quad (15)$$

where α_k is the cluster-specific intercept, $\beta_{k,1}, \beta_{k,2}$ measure the marginal sensitivities of fuel consumption to electric-mode share and annual driving distance, respectively, and ε_{ik} represents the random error (or residual) in the within-cluster regression model. This characterization enables an energy-oriented interpretation of the behavioral regimes determined by the GMM.

3. Model Validation and Evaluation Metrics

In this Section, the proposed model's reliability and behavioral interpretability are evaluated using quantitative validation metrics that measure cohesion, separation, and stability within the observed clusters. The first validation criterion is the Silhouette coefficient, which assesses how similar each observation is to its own cluster relative to other clusters [42]. For an observation i assigned to a cluster, the average intra-cluster distance $\xi(i)$ and the smallest mean inter-cluster distance $\zeta(i)$ are defined as

$$\xi(i) = \frac{1}{|C_k| - 1} \sum_{j \in C_k, j \neq i} \|x_i - x_j\|, \quad \zeta(i) = \min_{m \neq k} \frac{1}{|C_m|} \sum_{j \in C_m} \|x_i - x_j\|, \quad (16)$$

where $C_k = \{i \mid \hat{y}(i) = k\}$, $k = 1, 2, \dots, \hat{K}$, is the set of indices of all data points assigned to cluster k and $C_m = \{i \mid \hat{y}(i) = m\}$, is the set of observations belonging to cluster m (a different cluster than C_k). Then, the silhouette value for point i is given as

$$S(i) = \frac{\zeta(i) - \xi(i)}{\max\{\xi(i), \zeta(i)\}} \in [-1, 1], \quad (17)$$

where $S(i)$ close to 1 indicates that the observation is well matched to its own cluster and poorly matched to others, while values near 0 or negative denote boundary or misclassified points. The average silhouette over all observations gives a general measure of cluster compactness and separability.

To further evaluate the trade-off between intra-cluster compactness and inter-cluster separation, the Davies–Bouldin (DB) index is employed [43]. While the silhouette coefficient assesses pointwise cohesion, the DB index summarizes cohesion at the cluster level. Specifically, the silhouette coefficient relies on the average intra-cluster distance for each observation and evaluates how well each observation fits within its assigned cluster; the DB index summarizes this concept across the entire cluster using its mean intra-cluster dispersion. Hence, for each cluster C_k , the average distance of all points within that cluster to their centroid c_k can be defined as

$$R_k = \frac{1}{|C_k|} \sum_{i \in C_k} \|x_i - c_k\|, \quad (18)$$

where R_k represents the cluster's internal compactness or radius. The degree of separation between two clusters C_k and C_m is then given as

$$D_{km} = \frac{R_k + R_m}{\|c_k - c_m\|}, \quad (19)$$

which compares their combined dispersion with the Euclidean distance between their centroids. Then, the overall DB index is calculated as

$$DB = \frac{1}{K} \sum_{k=1}^K \max_{m \neq k} D_{km}. \quad (20)$$

A lower DB value indicates that clusters are compact and well-separated, whereas higher values imply overlapping or poorly defined cluster boundaries.

Furthermore, to quantify the robustness of the clustering outcome, a bootstrap stability analysis is conducted by refitting the model on resampled data and assessing label consistency using the Adjusted Rand Index (ARI) [44]. This analysis evaluates whether the identified clustering structure remains consistent when the observed data are resampled, thereby providing an additional measure of the robustness of unsupervised clustering results. The ARI quantifies the similarity between two cluster label assignments: one derived from the original dataset and the other from a resampled dataset, while accounting for the level of agreement anticipated by random chance. Let the two partitions be denoted as $\mathcal{C}_1$ and $\mathcal{C}_2$, each comprising K clusters. Their cross-classification is represented in a contingency matrix comprising various elements n_{rs} , where n_{rs} is the number of data points simultaneously assigned to cluster r in $\mathcal{C}_1$ and cluster s in $\mathcal{C}_2$. Then, the ARI is defined as

$$ARI = \frac{\sum_{r,s} \binom{n_{rs}}{2} - \frac{\sum_r \binom{a_r}{2} \sum_s \binom{b_s}{2}}{\binom{N}{2}}}{\frac{1}{2} \left[\sum_r \binom{a_r}{2} + \sum_s \binom{b_s}{2} \right] - \frac{\sum_r \binom{a_r}{2} \sum_s \binom{b_s}{2}}{\binom{N}{2}}}, \quad (21)$$

where $a_r = \sum_s n_{rs}$ and $b_s = \sum_r n_{rs}$ are the total number of points in respective clusters. An ARI value of 1 signifies complete consistency between the two clusters, whereas a value close to 0 denotes random agreement.

4. Numerical Experiments

4.1. Simulation Settings

The simulation experiments were designed to evaluate the performance, robustness, and interpretability of the GMM-based clustering framework under realistic PHEV usage patterns. The analysis used real-world data from 175 vehicle-level observations, each comprising six attributes: vehicle identifier, annual vehicle-kilometers traveled (VKT), utility factor (UF), number of observation days, and fuel consumption (L/100 km). Because UF is constrained to $[0, 1]$ and VKT exhibits a positively skewed distribution, both variables were transformed into a linear domain using (1). This transformation promotes homoscedasticity and enhances Gaussian separability in the latent behavioral space. In Figure 2, the standardized (z-score) values of the key variables are presented, including utility factor (UF), logarithm of annual vehicle kilometers traveled, and fuel consumption for all vehicles in the dataset. This visualization indicates substantial variation among vehicles and motivates the use of clustering to identify latent behavioral regimes.

The EM algorithm was employed for model estimation under the assumption of a full covariance, allowing each cluster to represent the anisotropic distribution of actual PHEV driving patterns. To avoid numerical degeneracy due to the moderate sample size, a covariance regularization term of 10^{-6} was employed to maintain positive definiteness of

Σ_k . The maximum iteration count and convergence tolerance were set to 2000 and 10^{-6} , respectively, to ensure numerical stability of the likelihood maximization. The optimal number of mixture components was determined by minimizing the BIC, which penalizes model complexity to prevent overfitting.

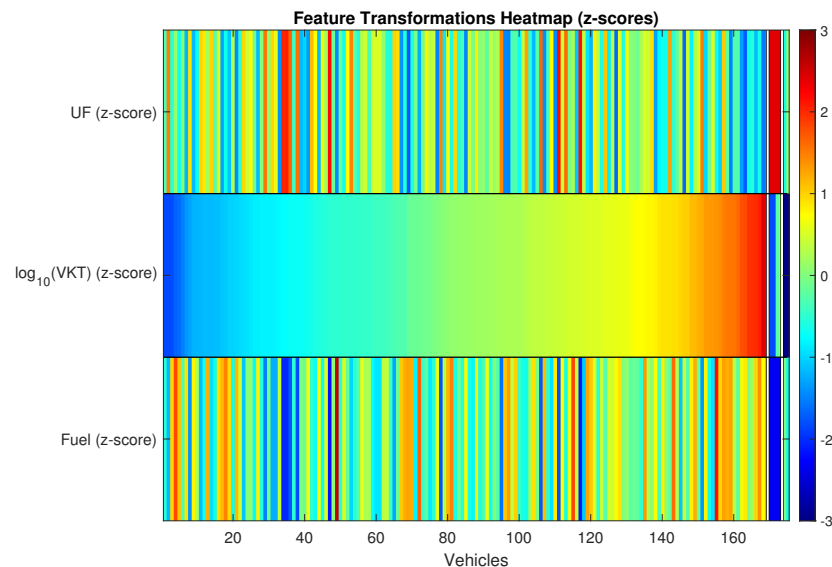


Figure 2. Feature transformation and standardization of behavioral variables.

All computations were performed in MATLAB R2024 on a 12-core Intel i7 workstation with 32 GB of RAM. Model quality was evaluated using multiple complementary indices: the mean Silhouette coefficient (cluster cohesion), the Davies–Bouldin (DB) index (cluster separation), and the Adjusted Rand Index (ARI) computed from 100 bootstrap resamples (model stability). For benchmarking, k-means [45] and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [46] clustering were applied to the same transformed data to test the sensitivity of the identified behavioral regimes to the clustering paradigm. The resulting clusters were visualized using UF–VKT scatter maps, Gaussian density contours, silhouette profiles, posterior entropy maps, and kernel-density overlays, providing a comprehensive interpretation of PHEV usage heterogeneity with respect to energy efficiency.

4.2. Results

4.2.1. Model Selection and Cluster Quality Evaluation

Following the simulation configuration described in Section 4.1, we first conducted the statistical evaluation of the clustering framework used to identify latent usage patterns in PHEVs. In Figure 3, three complementary aspects of model assessment are presented: model-order selection using the Bayesian Information Criterion (BIC), cluster cohesion measured by the mean Silhouette coefficient, and cluster separation quantified by the Davies–Bouldin (DB) index. Figure 3a illustrates the BIC values obtained from GMMs fitted with varying numbers of mixture components K . The BIC curve exhibits a pronounced decrease as model complexity increases from $K = 1$ to $K = 2$, followed by a minimum at $K = 3$. Beyond this point, the BIC increases again, indicating that additional mixture components introduce unnecessary model complexity without providing a significant improvement in likelihood. According to the BIC criterion, the optimal number of clusters is $\hat{K} = 3$, which represents the most parsimonious representation of the dataset’s underlying behavioral structure. This result suggests that the PHEV usage population can be reasonably described by three dominant operational regimes in the transformed UF–VKT feature space.

Although the detailed cluster-size distribution is presented later in Table 2, the selected three-component solution was further examined to ensure that the smaller components reflect meaningful behavioral structure rather than overfitting. As shown later by posterior uncertainty and Mahalanobis-distance analyses, these components occupy distinct regions of the transformed UF-VKT space. Therefore, they are interpreted cautiously as rare behavioral regimes rather than population-level subgroup estimates.



Figure 3. Model selection and clustering quality evaluation. (a) Bayesian Information Criterion (BIC) values for Gaussian mixture models with varying numbers of components, (b) comparison of mean silhouette scores between GMM and k-means clustering, and (c) Davies–Bouldin index comparison.

Table 2. Summary of PHEV Usage Clusters.

Cluster	Interpretation	Size (%)	Mean UF	Mean VKT (km/yr)	Fuel (L/100 km)
1	Medium EV usage	169 (96.6%)	0.380	18,617	3.89
2	High EV usage	4 (2.3%)	1.000	11,267	0.00
3	Low-mileage usage	2 (1.1%)	0.364	2444	3.16

Note: Clusters 2 and 3 represent rare behavioral patterns and therefore contain a small number of vehicles.

Figure 3b,c compare the clustering quality between the GMM and k-means algorithm using two widely adopted internal validation metrics. The mean Silhouette coefficient, shown in Figure 3b, evaluates the cohesion of clusters relative to their separation. The k-means method yields a higher Silhouette score (0.623) than the GMM (0.405) because the Silhouette metric inherently favors spherical cluster structures, which aligns with the assumptions of the k-means algorithm. While k-means yields a higher silhouette score due to its spherical cluster assumption, the probabilistic GMM provides a more flexible and statistically appropriate representation of the underlying data distribution. The DB index shown in Figure 3c assesses the ratio of intra-cluster dispersion to inter-cluster centroid distance. The GMM method yields a lower DB index (0.557) than k-means (0.589), indicating better cluster separation related to their internal dispersion. Alongside the k-means comparison, a density-based clustering method (DBSCAN) was applied to the same transformed feature space to assess the sensitivity of the identified behavioral regimes to the clustering paradigm. The DBSCAN identified four clusters with several noise points (63 observations). The resulting clustering exhibited a lower mean silhouette score (0.213) than both GMM (0.405) and k-means (0.623), indicating weaker cohesion and separation of clusters. The DB index (0.561) was comparable to that of the GMM (0.557) but higher than desirable for well-separated cluster structures. Hence, while DBSCAN can detect localized density structures, it is less effective in capturing the continuous behavioral regimes present in the UF–VKT feature space. The selected benchmark methods represent two common alternatives to model-based clustering: k-means as a centroid-based method and DBSCAN as a density-based method. Thus, they provide complementary comparisons for assessing cluster compactness, separation, and sensitivity to different clustering assumptions.

These comparisons highlight the advantage of the proposed GMM framework in capturing elliptical cluster geometries and correlated feature structures in the transformed behavioral space (u, v) . The combination of BIC-based model selection and an improved DB index supports adopting the GMM framework to identify latent PHEV usage regimes in the subsequent analysis.

4.2.2. Gaussian Mixture Clustering Structure and Assignment Uncertainty

Next, we assess the clustering structure delineated by the proposed GMM, along with the geometric characteristics of the clusters and the model's confidence in categorizing observations to behavioral regimes. In Figure 4, the cluster distribution in the original feature space, the probabilistic structure of the mixture components in the transformed space, and the uncertainty associated with cluster assignments are presented. Figure 4a shows the distribution of vehicles in the original UF–VKT behavioral space, where the annual vehicle-kilometers traveled (VKT) is shown on a logarithmic scale and the utility factor (UF) represents the proportion of electric driving. The scatter plot reveals three distinct clusters corresponding to the optimal number of components $\hat{K} = 3$ determined previously through BIC-based model selection. The majority of observations belong to Cluster 1, which occupies the central region of the UF–VKT domain and represents typical PHEV usage characterized by moderate annual mileage and intermediate electric-driving shares. Conversely, Clusters 2 and 3 comprise smaller groups of vehicles with more specialized usage patterns. Cluster 2 is associated with vehicles exhibiting high utility factors, indicating a strong reliance on electric propulsion, while Cluster 3 contains vehicles with lower annual driving distances and relatively lower UF values. The centroids marked with black crosses represent the mean behavioral profiles of each cluster, providing an intuitive summary of the dataset's dominant operating regimes.

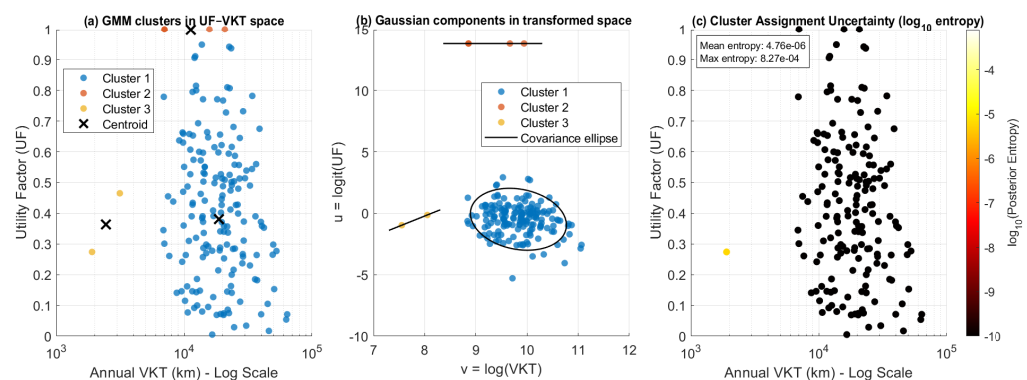


Figure 4. Gaussian mixture clustering results in the PHEV behavioral space. (a) Cluster distribution in the original UF–VKT space, (b) cluster structure in the transformed feature space, and (c) Posterior entropy of cluster assignments.

Figure 4b shows the clusters in the transformed feature space defined by (1). This transformation linearizes the bounded UF variable and stabilizes the variance of VKT, thereby enabling the GMM to better represent the data distribution. The covariance ellipse shows the shape and orientation of the principal Gaussian component associated with the dominant cluster. The elliptical shape indicates the correlated variability among the transformed variables, confirming that the clusters are not spherical but exhibit anisotropic dispersion. This feature demonstrates the superiority of the GMM framework over simpler clustering methods, such as k-means, which assume isotropic cluster shapes. Note that Cluster 2 and Cluster 3 appear as almost straight lines instead of ellipses. This is because small clusters exhibit elongated covariance structures due to limited sample size, resulting in nearly degenerate ellipses in the transformed space.

Figure 4c illustrates the uncertainty of cluster assignments using the posterior entropy measure derived from the GMM posterior probability. Each point is colored according to the logarithm of its entropy, providing a quantitative indication of classification confidence. The entropy values are very small across the dataset, with mean and maximum values on the order of 10^{-6} and 10^{-4} , respectively. Such low entropy levels indicate that the allocated cluster has a posterior probability close to one for almost all observations. In other words, the clusters identified by the GMM are highly distinct, and the model assigns vehicles to clusters with strong statistical confidence. These findings provide a strong empirical foundation for interpreting the clusters as distinct categories of PHEV operational behavior in subsequent analysis of energy use.

4.2.3. Cluster-Specific Behavioral and Energy Consumption Characteristics

In Figure 5, the behavioral and energy-related characteristics associated with each cluster are presented, building upon the clustering structure identified in the previous subsection. Specifically, we analyze the distinctions among these clusters in their underlying distributions of utility factor, driving intensity, and fuel consumption to understand PHEVs' energy-using behavior.

Figure 5a shows that Cluster 2 exhibits the highest UF values, approaching unity, signifying that vehicles in this cluster operate almost entirely in electric mode. Conversely, Cluster 1 exhibits a broader UF distribution with a median in the mid-range, indicating a mixed-mode operation that utilizes both electric and combustion propulsion. Cluster 3 exhibits moderately lower UF values and a narrower distribution, indicating a distinct usage group characterized by a relatively limited electric driving share. Figure 5b illustrates the distribution of annual VKT across clusters using a logarithmic transformation. Cluster 1, containing the majority of observations, represents typical PHEV users with moderate to high annual mileage. Cluster 2 exhibits comparable or slightly lower mileage while retaining distinctive characteristics due to its very high electric usage. Cluster 3 is characterized by noticeably lower annual driving distances, indicating a small group of low-mileage users. These patterns reinforce the behavioral interpretation of the previously identified clusters: the dominant cluster represents conventional mixed-mode usage, while the remaining clusters capture specialized driving regimes. Figure 5c further highlights the energy implications of these behavioral differences by comparing fuel consumption across clusters. Vehicles in Cluster 2 exhibit near-zero fuel consumption, which is consistent with the very high utility factor observed in Figure 5a, which indicates that these vehicles rely almost exclusively on electric propulsion during the observation period. Cluster 1 exhibits the highest variability in fuel consumption, reflecting a wide range of mixed electric–combustion driving patterns. In contrast, Cluster 3 shows moderate fuel consumption, consistent with lower driving intensity and a smaller share of electric operation.

Figure 5d illustrates the relative sizes of the clusters, indicating a highly imbalanced distribution in which Cluster 1 contains approximately 96.6% of the observations, whilst Clusters 2 and 3 represent only small fractions of the dataset. Although these minor clusters represent a small number of vehicles, they capture important niche behaviors, such as highly electrified driving patterns and low-mileage usage regimes that would be difficult to identify without a probabilistic clustering framework. This highlights the heterogeneity of real-world PHEV usage patterns even within relatively small datasets. Figure 5e shows a clear negative relationship between UF and fuel consumption, indicating that higher electric driving shares are associated with lower gasoline usage. This pattern aligns with the operating principles of PHEVs and provides an intuitive validation of the clustering results. Figure 5f depicts the relationship between annual driving distance and fuel consumption. Although the correlation is less evident than in Figure 5e, vehicles

with higher VKT generally demonstrate greater variability in fuel consumption owing to variations in driving behavior and electric charging availability. The combination of distributional comparisons and relationships among variables confirms that the proposed GMM framework effectively captures variations in both driving patterns and energy usage. The cluster interpretation summary is given in Table 2.

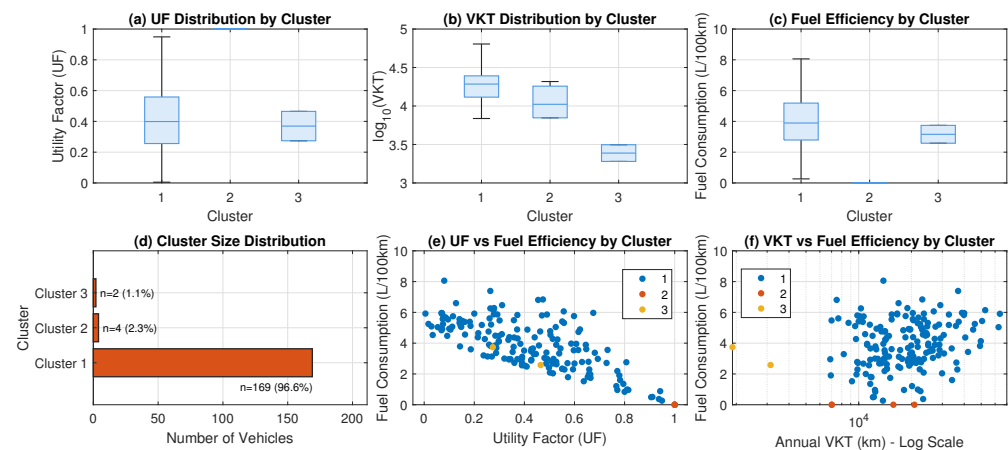


Figure 5. Cluster-specific characteristics of behavioral and fuel consumption. (a) UF distribution by clusters, (b) VKT distribution by clusters, (c) fuel efficiency by clusters, (d) cluster size distribution, (e) UF vs fuel efficiency by cluster, and (f) VKT vs fuel efficiency by cluster.

4.2.4. Clustering Diagnostics and Model Robustness Assessment

Following the interpretation of cluster-specific behavioral and energy characteristics, it is important to evaluate the robustness and statistical validity of the identified clustering structure. In Figure 6, the diagnostic metrics used to assess the stability, confidence, and geometric separation of the GMM clusters are presented. Figure 6a shows the bootstrap stability analysis using the Adjusted Rand Index (ARI), which measures the agreement between clustering assignments obtained from bootstrap resampled datasets and the original dataset. The histogram depicts the distribution of ARI values from repeated bootstrap iterations, resulting in a mean ARI of approximately 0.448. This value indicates moderate clustering stability, which is expected given the highly imbalanced cluster structure previously observed, in which the majority of observations belong to Cluster 1, whereas Clusters 2 and 3 represent relatively small groups. In such situations, bootstrap resampling can slightly perturb the composition of smaller clusters, leading to variations in label assignments. Nonetheless, the ARI distribution remains consistently above random agreement, indicating that the clustering solution reflects meaningful structure rather than artifacts of sampling variability. Nevertheless, because Clusters 2 and 3 contain only a few observations, their stability should be interpreted cautiously and further validated using larger datasets.

Figure 6a illustrates the posterior probability matrix obtained from the GMM. Each column corresponds to an observation, while each row represents the posterior probability of membership in a particular cluster. The matrix exhibits a nearly one-hot structure, in which most observations have a probability close to one for a single cluster and near-zero probabilities for the others. This pattern indicates that the model assigns vehicles to clusters with high confidence, reinforcing the interpretation from the previous section that the clusters represent distinct behavioral regimes. Only a small number of observations show slightly distributed probabilities, which may correspond to vehicles with transitional driving behaviors between clusters. Figure 6c further assesses the geometric separation of clusters by displaying the pairwise Mahalanobis distances between cluster centroids in the transformed feature space. The Mahalanobis distance captures the covariance struc-

ture of the data, providing a scale-invariant measure for distinguishing between clusters. The findings indicate substantial separation among cluster centroids, especially between Clusters 2 and 3, which exhibit the largest distance. Cluster 1 is also clearly separated from the other two clusters. The distances confirm that the clusters inhabit distinct regions of the feature space, hence validating the GMM-based segmentation. Notably, despite the inherent imbalance in cluster sizes, the GMM demonstrates high assignment confidence and clear geometric separation between clusters. Therefore, combined with the behavioral and energy-use interpretations, these diagnostics strengthen the overall reliability of the clustering framework and support its use for characterizing heterogeneous PHEV usage patterns. The above findings provide a strong foundation for PHEV policy implications as discussed in the next section.

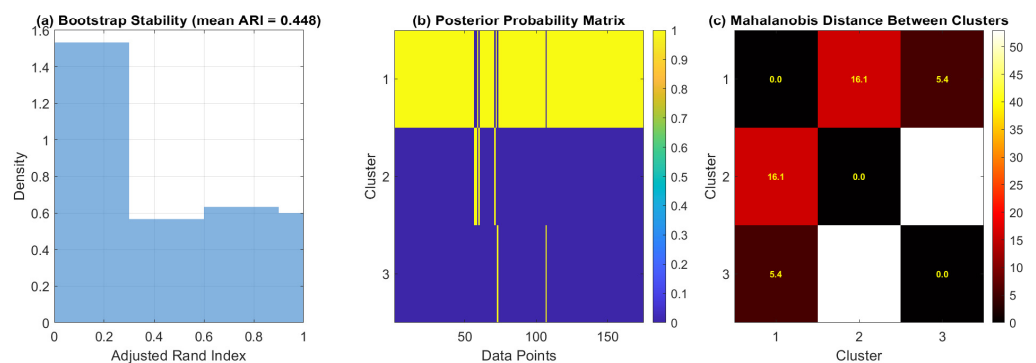


Figure 6. Clustering diagnostics and validation of the GMM. (a) Bootstrap stability analysis of the clustering results measured using the ARI, (b) Posterior probability matrix showing the cluster membership probabilities for each observation, and (c) Pairwise Mahalanobis distances between cluster centroids in the transformed feature space, illustrating the degree of separation between clusters.

4.2.5. Policy Implications of PHEV Usage Patterns

The clustering results reveal clear heterogeneity in real-world PHEV operational behavior, which has important implications for charging infrastructure planning and policy design. The dominant cluster, representing approximately 96.6% of vehicles, exhibits moderate electric usage (mean UF ≈ 0.38) alongside relatively high annual mileage ($\approx 18,600$ km), suggesting that many PHEV users operate in mixed electric-combustion modes, likely due to limited charging opportunities during daily travel. This finding implies that the strategic expansion of workplace and public charging infrastructure could substantially increase the share of electric driving within this large user group. Conversely, the smaller cluster characterized by nearly full electric operation (UF ≈ 1.0) highlights that reliable charging access enables PHEVs to function effectively as electric vehicles in real-world conditions. These findings highlight the importance of strategically locating charging infrastructure in urban areas and along common commuting corridors to maximize electric-mode utilization across the broader vehicle fleet.

The identified behavioral regimes also provide insights for targeted policy incentives and adaptive energy management strategies. For example, drivers in the dominant mixed-usage cluster represent the greatest opportunity to reduce fuel consumption through behavioral or technological interventions, such as incentive programs that encourage regular charging or smart energy management systems that prioritize electric operation when battery charge is available. Meanwhile, the low-mileage cluster indicates a small subset of vehicles with limited annual driving demand, suggesting that PHEVs may serve different roles within household mobility portfolios. From a policy perspective, differentiating between urban users with shorter travel distances and high-UF potential,

and commuter-oriented users with higher VKT, could enable more efficient deployment of incentives and infrastructure investments.

Although the present analysis does not directly quantify emission reduction or charging-infrastructure impacts, the identified clusters provide a practical basis for such future assessments. In particular, the dominant mixed-mode cluster, which combines moderate UF with relatively high annual mileage, represents the group with the greatest potential for fuel reduction if additional charging opportunities increase electric-mode use. By contrast, the highly electrified cluster illustrates the upper bound of PHEV electric operation under favorable charging conditions, while the low-mileage cluster is expected to have a smaller aggregate impact due to its limited annual travel demand. Therefore, the proposed clustering framework can serve as an intermediate step between vehicle-level usage characterization and future quantitative models of charging demand, fuel-saving potential, and emission reduction. Such quantitative assessment would require additional information on charging locations, charging frequency, trip-level driving patterns, electricity carbon intensity, and local infrastructure availability, which are beyond the scope of the present dataset.

From a systems perspective, the identified behavioral heterogeneity has direct implications for aggregate energy demand, charging load distribution, and transportation network efficiency. The dominance of mixed-mode users suggests that current infrastructure and policy configurations do not fully enable electric-mode utilization at the fleet level, leading to suboptimal system-wide energy performance. In contrast, the presence of fully electrified users demonstrates the achievable upper bound of electrification under adequate charging availability. These findings highlight that system-level optimization of PHEV benefits requires coordinated design of infrastructure, policy incentives, and energy management strategies to align individual behavior with network-wide sustainability objectives.

5. Conclusions

This study presented a data-driven framework for identifying and characterizing heterogeneous real-world usage patterns of PHEVs. Using UF and VKT as the two primary behavioral indicators, the proposed approach applied logit and logarithmic transformations to obtain a statistically suitable feature space for Gaussian mixture modeling. The full-covariance GMM, combined with BIC-based model selection, identified three PHEV operational regimes: a dominant mixed electric-combustion usage group, a small highly electrified group, and a small low-mileage group.

The results show that the dominant cluster, representing the majority of vehicles, exhibits moderate electric-mode utilization and relatively high annual mileage, indicating that many PHEVs are still operated in mixed electric-combustion mode. This group therefore represents an important opportunity for improving electric-mode usage through better charging access, targeted incentives, and adaptive energy-management strategies. The highly electrified cluster demonstrates that near-electric operation is achievable under favorable usage and charging conditions, while the low-mileage cluster reflects a distinct group with limited annual mobility demand. The post-clustering analysis further confirms that higher UF is associated with lower fuel consumption, supporting the behavioral and energy-related interpretation of the identified clusters.

The robustness analysis indicates that the GMM yields high posterior-assignment confidence and clear geometric separation among the identified regimes. However, the bootstrap ARI indicates moderate stability, mainly due to the highly imbalanced cluster structure and the small number of observations in the two rare clusters. Therefore, these small clusters should be interpreted with caution as indicative behavioral patterns rather

than as statistically representative population-level groups. This limitation highlights the need for validation on larger, more diverse PHEV datasets.

From a policy and systems perspective, the proposed framework can support differentiated charging infrastructure planning, user-specific incentives, and more effective PHEV energy management strategies. In particular, identifying mixed-mode users with high annual mileage may help prioritize interventions that can yield greater fuel-saving and emission-reduction benefits. Future work could extend this approach toward multi-scale system analysis, linking vehicle-level behavior with network-level traffic dynamics and grid-level energy demand to support the design of fully integrated sustainable transport systems. Incorporating additional behavioral variables such as charging frequency, ambient temperature, battery capacity, trip-level driving patterns, and temporal driving dynamics could be interesting future work. Moreover, integrating larger and more diverse datasets, including regional or longitudinal vehicle usage records, would enable more comprehensive characterization of PHEV operational behavior across different transportation contexts. These studies could facilitate the development of adaptive vehicle energy management strategies and data-driven transportation policies to maximize the environmental benefits of electrified mobility.

Author Contributions: Conceptualization, A.S.M.B.; methodology, A.S.M.B., M.A.S.K. and J.-i.I.; software, A.S.M.B.; validation, A.S.M.B. and M.A.S.K.; formal analysis, A.S.M.B.; investigation, A.S.M.B.; resources, A.S.M.B.; data curation, A.S.M.B.; writing—original draft preparation, A.S.M.B.; writing—review and editing, M.A.S.K. and J.-i.I.; visualization, A.S.M.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data are available at <https://data.mendeley.com/datasets/vmfsh6rg5t/1> (accessed on 15 December 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Bakibillah, A.S.M.; Kamal, M.A.S.; Tan, C.P.; Hayakawa, T.; Imura, J.I. Event-driven stochastic eco-driving strategy at signalized intersections from self-driving data. *IEEE Trans. Veh. Technol.* **2019**, *68*, 8557–8569.
2. U.S. Energy Information Administration (EIA). *Annual Energy Outlook 2019: With Projections to 2050*; U.S. Department of Energy: Washington, DC, USA, 2019.
3. U.S. Environmental Protection Agency (EPA). *Inventory of U.S. Greenhouse Gas Emissions and Sinks: 1990–2017*; Environmental Protection Agency: Washington, DC, USA, 2019.
4. Pietrzak, K.; Pietrzak, O. Environmental effects of electromobility in a sustainable urban public transport. *Sustainability* **2020**, *12*, 1052. [\[CrossRef\]](#)
5. Jayson, T.; Bakibillah, A.S.M.; Tan, C.P.; Kamal, M.A.S.; Monn, V.; Imura, J.I. Electric vehicle eco-driving strategy at signalized intersections based on optimal energy consumption. *J. Environ. Manag.* **2024**, *368*, 122245. [\[CrossRef\]](#)
6. Ding, F.; Luo, X.; Li, G.; Tew, H.H.; Loo, J.Y.; Tong, C.W.; Bakibillah, A.S.M.; Zhao, Z.; Tao, Z. Energy-efficient hybrid model predictive trajectory planning for autonomous electric vehicles. In Proceedings of the 2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Sarawak, Malaysia, 6–9 October 2024; pp. 1263–1269.
7. Marotta, A.; Lodi, C.; Julea, A.; Gómez Vilchez, J.J. European governments' electromobility plans: An assessment with a focus on infrastructure targets and vehicle estimates until 2030. *Energy Effic.* **2023**, *16*, 92. [\[CrossRef\]](#)
8. Lopez-Arboleda, E.; Sarmiento, A.T.; Cardenas, L.M. Policy assessment for electromobility promotion in Colombia: A system dynamics approach. *Transp. Res. Part D Transp. Environ.* **2023**, *121*, 103799. [\[CrossRef\]](#)
9. Xiong, R.; Cao, J.; Yu, Q. Reinforcement learning-based real-time power management for hybrid energy storage system in the plug-in hybrid electric vehicle. *Appl. Energy* **2018**, *211*, 538–548.

10. Lei, Z.; Qin, D.; Hou, L.; Peng, J.; Liu, Y.; Chen, Z. An adaptive equivalent consumption minimization strategy for plug-in hybrid electric vehicles based on traffic information. *Energy* **2020**, *190*, 116409. [\[CrossRef\]](#)
11. Qian, Y. The Impact of the Bursting New Energy Vehicle Market on the Traditional Fuel Vehicle Industry—The Case of Audi. In *Proceedings of the Advances in Economics, Management and Political Sciences*; Eliwise Academy: Oxford, UK, 2024; Volume 91, pp. 237–242.
12. Şenyapar, H.N.D.; Tür, M.R. Turkey's Domestic Car TOGG in the Framework of New Trends in Electric Vehicles and Consumer Ethnocentrism. *Balk. J. Electr. Comput. Eng.* **2023**, *11*, 115–128. [\[CrossRef\]](#)
13. Joshi, P.; Powell, B.; Weigl, D.; Johnson, C.; Man, D. *Advancing Transportation Efficiency and Electric Vehicles in Tonga: A Review of Relevant Trends and Best Practices*; Technical Report NREL/TP-7A40-84518; National Renewable Energy Laboratory (NREL): Golden, CO, USA, 2023.
14. Plötz, P.; Funke, S.Á.; Jochem, P. Empirical fuel consumption and CO₂ emissions of plug-in hybrid electric vehicles. *J. Ind. Ecol.* **2018**, *22*, 773–784.
15. Jeong, J.W.; Lee, G.; Lee, J.; Woo, S.; Kim, N.; Lee, K. Analysis of energy consumption and emissions characteristics of plug-in hybrid electric vehicle (PHEV) under various real-world driving conditions. *J. Environ. Sci.* **2025**, *163*, 566–574.
16. Suarez-Bertoa, R.; Pavlovic, J.; Trentadue, G.; Otura-Garcia, M.; Tansini, A.; Ciuffo, B.; Astorga, C. Effect of low ambient temperature on emissions and electric range of plug-in hybrid electric vehicles. *ACS Omega* **2019**, *4*, 3159–3168. [\[CrossRef\]](#) [\[PubMed\]](#)
17. Plötz, P.; Moll, C.; Bieker, G.; Mock, P. From lab-to-road: Real-world fuel consumption and CO₂ emissions of plug-in hybrid electric vehicles. *Environ. Res. Lett.* **2021**, *16*, 054078.
18. Zhou, B.; Zhang, S.; Wu, Y.; Ke, W.; He, X.; Hao, J. Energy-saving benefits from plug-in hybrid electric vehicles: Perspectives based on real-world measurements. *Mitig. Adapt. Strateg. Glob. Change* **2018**, *23*, 735–756.
19. Frey, H.C.; Zheng, X.; Hu, J. Variability in measured real-world operational energy use and emission rates of a plug-in hybrid electric vehicle. *Energies* **2020**, *13*, 1140.
20. Graver, B.M.; Frey, H.C.; Choi, H.W. In-use measurement of activity, energy use, and emissions of a plug-in hybrid electric vehicle. *Environ. Sci. Technol.* **2011**, *45*, 9044–9051. [\[CrossRef\]](#) [\[PubMed\]](#)
21. Mandev, A.; Plötz, P.; Sprei, F. Factors impacting real-world fuel economy of plug-in hybrid electric vehicles in Europe—An empirical analysis. *Environ. Res. Commun.* **2024**, *6*, 051001.
22. Graba, M.; Hennek, K.; Mamala, J.; Bieniek, A.; Burdzik, R. Factors Influencing Energy Consumption in Plug-In Hybrid Vehicles—A Field Study. *SSRN Electron. J.* **2025**. [\[CrossRef\]](#)
23. Williams, B.; Martin, E.; Lipman, T.; Kammen, D. Plug-in-hybrid vehicle use, energy consumption, and greenhouse emissions: An analysis of household vehicle placements in northern California. *Energies* **2011**, *4*, 435–457.
24. Hu, X.; Moura, S.J.; Murgovski, N.; Egardt, B.; Cao, D. Integrated optimization of battery sizing, charging, and power management in plug-in hybrid electric vehicles. *IEEE Trans. Control Syst. Technol.* **2015**, *24*, 1036–1043. [\[CrossRef\]](#)
25. Xie, S.; Qi, S.; Lang, K.; Tang, X.; Lin, X. Coordinated management of connected plug-in hybrid electric buses for energy saving, inter-vehicle safety, and battery health. *Appl. Energy* **2020**, *268*, 115028.
26. Li, Y.; Wang, F.; Tang, X.; Hu, X.; Lin, X. Convex optimization-based predictive and bi-level energy management for plug-in hybrid electric vehicles. *Energy* **2022**, *257*, 124672.
27. Sun, X.; Fu, J.; Yang, H.; Xie, M.; Liu, J. An energy management strategy for plug-in hybrid electric vehicles based on deep learning and improved model predictive control. *Energy* **2023**, *269*, 126772. [\[CrossRef\]](#)
28. Han, Q.; Zhao, X.; Ma, J.; Wang, S.; Zhang, K.; Chai, H.; Yang, Y.; Zheng, Z. Transfer Deep Reinforcement Learning-Based Energy Management Strategy for Plug-In Hybrid Electric Heavy-Duty Trucks under Segmented Usage Scenarios. *Int. J. Energy Res.* **2023**, *2023*, 1875380.
29. Zhang, X.; Jiao, X.; Zhang, Y.; Li, Z. Adaptive Weight-Based Predictive Energy Management Strategy for Plug-in Hybrid Electric Vehicles Considering Battery Life. In *Proceedings of the 2025 9th CAA International Conference on Vehicular Control and Intelligence (CVCI)*, Changsha, China, 24–26 October 2025; pp. 1–6.
30. Dang, H.; Han, Y.; Hao, Y.; Sun, P.; Chen, Z. Energy management optimization of plug-in hybrid electric vehicle in microgrid with information-physics-traffic coupling. *Electr. Power Syst. Res.* **2025**, *238*, 111194.
31. Doikin, A.; Korsunovs, A.; Campean, F.; García-Afonso, O.; Agostinelli, E. ML-Based Control Strategy for PHEV Under Predictive Vehicle Usage Behaviour. *Vehicles* **2025**, *7*, 23. [\[CrossRef\]](#)
32. Eissa, M.A.; Chen, P. Personalized electric vehicle range prediction based on self-supervised driving pattern clustering. In *Proceedings of the 2023 IEEE International Automated Vehicle Validation Conference (IAVVC)*, Austin, TX, USA, 16–18 October 2023; pp. 1–6.
33. Van Kriekinge, G.; De Cauwer, C.; Sapountzoglou, N.; Coosemans, T.; Messagie, M. Electric vehicle charging sessions generator based on clustered driver behaviors. *World Electr. Veh. J.* **2023**, *14*, 37. [\[CrossRef\]](#)
34. Plötz, P.; Funke, S.Á.; Jochem, P. The impact of daily and annual driving on fuel economy and CO₂ emissions of plug-in hybrid electric vehicles. *Transp. Res. Part A Policy Pract.* **2018**, *118*, 331–340. [\[CrossRef\]](#)

35. Chen, X.; Zhang, C.; Cheng, Z.; Hou, Y.; Sun, L. A Bayesian Gaussian Mixture Model for probabilistic modeling of car-following behaviors. *IEEE Trans. Intell. Transp. Syst.* **2023**, *25*, 5880–5891. [[CrossRef](#)]
36. Sammaknejad, N.; Zhao, Y.; Huang, B. A review of the expectation maximization algorithm in data-driven process identification. *J. Process Control* **2019**, *73*, 123–136. [[CrossRef](#)]
37. Jannah, W.; Saputro, D.R. Parameter estimation of Gaussian mixture models (GMM) with expectation maximization (EM) algorithm. In *Proceedings of the AIP Conference Proceedings*; AIP Publishing LLC: Melville, NY, USA, 2022; Volume 2566, p. 040002.
38. Mouret, F.; Hippert-Ferrer, A.; Pascal, F.; Tourneret, J.Y. A robust and flexible EM algorithm for mixtures of elliptical distributions with missing data. *IEEE Trans. Signal Process.* **2023**, *71*, 1669–1682. [[CrossRef](#)]
39. Teklehaymanot, F.K.; Muma, M.; Zoubir, A.M. Bayesian cluster enumeration criterion for unsupervised learning. *IEEE Trans. Signal Process.* **2018**, *66*, 5392–5406. [[CrossRef](#)]
40. Zhang, J.; Yang, Y.; Ding, J. Information criteria for model selection. *Wiley Interdiscip. Rev. Comput. Stat.* **2023**, *15*, e1607. [[CrossRef](#)]
41. Jiang, M.; Xiang, Y.; Sheng, C. Based on spatial Mahalanobis distance: A novel zero-shot learning method for compound fault identification and decoupling. *Expert Syst. Appl.* **2025**, *272*, 126680. [[CrossRef](#)]
42. Lai, H.; Huang, T.; Lu, B.; Zhang, S.; Xiaog, R. Silhouette coefficient-based weighting k-means algorithm. *Neural Comput. Appl.* **2025**, *37*, 3061–3075.
43. Ying, R.; Taniguchi, T.; Nabeta, K.; Ishigami, K.; Kikuchi, N.; Okamoto, S. Clustering First-order reversal curve diagram using the Gaussian mixture model and the Davies–Bouldin index. *Jpn. J. Appl. Phys.* **2024**, *63*, 11SP04.
44. Sundqvist, M.; Chiquet, J.; Rigail, G. Adjusting the adjusted rand index: A multinomial story. *Comput. Stat.* **2023**, *38*, 327–347.
45. Sinaga, K.P.; Yang, M.S. Unsupervised K-means clustering algorithm. *IEEE Access* **2020**, *8*, 80716–80727. [[CrossRef](#)]
46. Hanafi, N.; Saadatfar, H. A fast DBSCAN algorithm for big data based on efficient density calculation. *Expert Syst. Appl.* **2022**, *203*, 117501. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.