


Article

Vehicular Fuel Consumption and CO₂ Emission Estimation Model Integrating Novel Driving Behavior Data Using Machine Learning

Ziyang Wang ^{1,*} , Masahiro Mae ¹, Shoma Nishimura ² and Ryuji Matsuhashi ¹

¹ Department of Electrical Engineering and Information Systems, The University of Tokyo, Tokyo 113-8656, Japan; mmae@ieee.org (M.M.); matu@enesys.t.u-tokyo.ac.jp (R.M.)

² Department of Digital Business Design, Aioi Nissay Dowa Insurance Co., Ltd., Tokyo 150-8488, Japan; shoma-nishimura@aioinissaydowa.co.jp

* Correspondence: wang-ziyang@ieee.org; Tel.: +86-156-5291-1197

Abstract: Fossil fuel vehicles significantly contribute to CO₂ emissions due to their high consumption of fossil fuels. Accurate estimation of vehicular fuel consumption and the associated CO₂ emissions is crucial for mitigating these emissions. Although driving behavior is a vital factor influencing fuel consumption and CO₂ emissions, it remains largely unaddressed in current CO₂ emission estimation models. This study incorporates novel driving behavior data, specifically counts of occurrences of dangerous driving behaviors, including speeding, sudden accelerating, and sudden braking, as well as driving time and driving distances on expressways, national highways, and local roads, respectively, into monthly fuel consumption estimation models for individual gasoline and hybrid vehicles. The CO₂ emissions are then calculated as a secondary parameter based on the estimated fuel consumption, assuming a linear relationship between the two. Using regression algorithms, it has been demonstrated that all the proposed driving behavior data are relevant for monthly CO₂ emission estimation. By integrating the driving behavior data of various vehicle categories, a generalizable CO₂ estimation model is proposed. When utilizing all the proposed driving behavior data collectively, our random forest regression model achieves the highest prediction accuracy, with R², RMSE, and MAE values of 0.975, 13.293 kg, and 8.329 kg, respectively, for monthly CO₂ emission estimation of individual vehicles. This research offers insights into CO₂ emission reduction and energy conservation in the road transportation sector.

Keywords: vehicular CO₂ emission; eco-driving; dangerous driving behavior; machine learning; random forest



Citation: Wang, Z.; Mae, M.; Nishimura, S.; Matsuhashi, R. Vehicular Fuel Consumption and CO₂ Emission Estimation Model Integrating Novel Driving Behavior Data Using Machine Learning. *Energies* **2024**, *17*, 1410. <https://doi.org/10.3390/en17061410>

Academic Editor: Pavel A. Strizhak

Received: 19 February 2024

Revised: 12 March 2024

Accepted: 13 March 2024

Published: 14 March 2024



Copyright: © 2024 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) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In 2020, road transportation accounted for 24% of the European Union's total carbon dioxide (CO₂) emissions, making it as the predominant greenhouse gas responsible for global warming [1]. Data from 2021 reveal that China's transportation sector consumed energy equivalent to approximately 530 million tons of standard coal, accounting for 15.3% of the country's overall energy consumption [2]. In the same vein, emissions from all passenger vehicles in Japan in 2020 made up 8.9% of the country's total CO₂ emissions [3]. By 2022, the International Energy Agency (IEA) reported that personal vehicles, such as cars and vans, were responsible for more than a quarter of the global oil consumption and approximately 10% of the worldwide CO₂ emissions linked to energy use [4].

The adverse effects of CO₂ emissions on both environmental sustainability and public health have been thoroughly documented in various studies, emphasizing the urgent need to substantially reduce these emissions [5]. This urgency is driven primarily by two reasons: First, the global commitment to mitigate climate change impacts, and second, the understanding that reducing CO₂ emissions, through measures like decreasing fuel

consumption, contributes to conserving our limited fossil fuel reserves [6,7]. However, despite the adoption of various strategies encompassing technological innovations [8] and policy regulations to curb CO₂ emissions, the increasing number of gasoline vehicles worldwide still presents a significant challenge to CO₂ emission reduction efforts [9–12], emphasizing the importance of accurate vehicular CO₂ emission estimation.

Accurate vehicular CO₂ emission estimation is crucial, as it not only provides feedback to drivers but also bolsters efforts in CO₂ emission reduction and energy conservation. In response to the mounting concerns over CO₂ emissions, Japan introduced the J-Credit Scheme in 2013 [13]. This government initiative promotes activities that diminish greenhouse gas emissions, aligning with the objectives of the Paris Agreement [14]. The J-Credit Scheme issues “J-Credits” for a myriad of activities, like renewable energy projects, energy efficiency enhancements, or activities that absorb CO₂, such as afforestation. Once these credits are issued, they can be traded, enabling entities to count them towards their greenhouse gas reduction targets. Such a mechanism bolsters the financing and incentivization of greenhouse gas reduction endeavors. There is potential in integrating CO₂ emission estimation results with the J-Credit Scheme, enhancing its effectiveness.

Prominent vehicular CO₂ emission models like MOBILE6 [15], developed by the U.S. Environmental Protection Agency (EPA), and EMFAC7F [16], proposed by the California Air Resources Board (CARB), strive to integrate factors related to travel, weather, and vehicle characteristics in their CO₂ emission estimations. In particular, these models primarily utilize metrics such as average speed and total vehicle miles traveled for CO₂ emission calculations. However, they often overlook the pivotal influences of roadway conditions, traffic dynamics, and drivers’ behaviors on vehicular CO₂ emissions. Several studies have explored this area in depth. In [17], Oduro et al. introduced a dynamic real-time fuel consumption estimation model based on multiple linear regression (MLR), underscoring a linear relationship between CO₂ emissions and both vehicle speed and acceleration. This suggests that these two factors play a significant role in determining CO₂ emission levels. In [18], Ahn et al. presented microscopic fuel consumption and emission models that leverage instantaneous speed and acceleration data by utilizing polynomial and hybrid regression models. This approach offers granular insights into the microscopic interactions between driving behaviors and fuel usage. In our preliminary work [19], we employed the random forest algorithm to estimate instantaneous fuel consumption using the instantaneous speed and acceleration data of a specific vehicle type in Japan. While estimating instantaneous fuel consumption has its merits, such as providing immediate feedback for efficient driving, promoting eco-driving, and aiding in traffic planning, it has limitations in capturing the broader picture. The focus on instantaneous CO₂ emissions can sometimes overshadow the significance of long-term accumulated CO₂ emissions, which are often of greater concern.

It has been revealed that the higher the frequency of speeding, sudden acceleration, and sudden braking, the lower the efficiency of the internal combustion engine [20,21]. In [22,23], Lois et al. and Jimenez et al. demonstrated that driving behaviors such as the rate of deceleration, revolutions per minute (RPM), and speed significantly influence fuel consumption. In [24], Mane et al. indicated that factors such as average speed, braking, and idling significantly influence fuel consumption. Congestion exacerbates this issue, leading to more frequent braking and reduced fuel efficiency, as reported in [25,26]. In contrast, Samaras et al. [27] demonstrated that free flow conditions result in a 4.9% decrease in fuel usage for a standard Euro 5 diesel vehicle with an engine capacity of less than 1.41 compared to congested scenarios. Moreover, Zhang et al. [28] reported a substantial difference in fuel consumption, exceeding 20%, between ecological and aggressive driving behaviors. Nevertheless, it is noteworthy that a gap still persists in existing studies that directly integrates driving behaviors into CO₂ emission estimation.

To tackle the urgent environmental concerns posed by increasing CO₂ emissions from vehicles, this paper introduces a novel methodology aimed at estimating monthly fuel consumption (V_{fuel}) for individual vehicles by leveraging detailed driving behavior data

provided by Aioi Nissay Dowa Insurance Co., Ltd. (Tokyo, Japan), which encompasses counts of risky driving actions, such as speeding, sudden acceleration, and sudden braking, across various road types, grounded in machine learning regression techniques. The monthly CO₂ emissions (E_{CO_2}) are then calculated as a secondary parameter based on the estimated fuel consumption, assuming a linear relationship between the two.

The significance of this work lies in how it leverages data on dangerous driving behaviors to enhance the accuracy of fuel consumption estimates. Unlike previous studies, which primarily focus on vehicle technical specifications and average driving patterns, our research offers a more nuanced understanding of the relationship between driving behaviors and fuel consumption. By providing a detailed analysis that connects individual driving actions to fuel consumption, this study not only enriches the existing body of knowledge on fuel consumption estimation models but also introduces a practical tool for drivers to evaluate and improve their driving habits towards more fuel-efficient practices. The CO₂ emissions are then estimated based on the fuel consumption, assuming a linear relationship between the two. It is important to note that the E_{CO_2} values were not directly verified by measurements, but rather estimated based on this assumed relationship. Moreover, the high prediction accuracy of the proposed fuel consumption estimation model and the associated CO₂ emission model holds promise for integration with initiatives like the J-Credit Scheme, potentially serving as a catalyst for the widespread adoption of fuel-efficient driving behaviors through targeted incentives. By aligning individual drivers' interests with broader environmental objectives, our approach offers a novel pathway to mitigate vehicular fuel consumption and the associated CO₂ emissions. The remainder of this paper is organized as follows: Section 2 details the methodology of this study. Section 3 shows the overall structure of the estimation models for V_{fuel} and E_{CO_2} , Section 4 discusses the performance and implications of the prediction results. Finally, Section 5 concludes the paper with a summary of our contributions and suggestions for future research.

2. Methodology

In this study, we investigate the potential of driving behavior data as features for estimating E_{CO_2} . Then, multiple feature sets are defined to evaluate the proposed driving characteristics. Subsequently, dimensionality reduction algorithms and correlation analysis are employed on the driving characteristics for data visualization. Lastly, machine learning regressions are conducted to evaluate the performance of each feature set in estimating E_{CO_2} .

2.1. Novel Driving Behavior Data

We base our investigation on 228,281 monthly driving behavior data instances from medium-sized TOYOTA vehicles (Toyota: Toyota City, Japan), provided by Aioi Nissay Dowa Insurance Co., Ltd. Table 1 details the statistics of the categories (model, vehicle name, and engine type) of the vehicles. For confidentiality reasons and for the purposes of both dimensionality reduction and machine learning analysis, the actual names of the model, vehicle name, and engine type are encoded into numerical labels 1, 2, 3, ..., as illustrated in Table 1. For the engine type, the numerical label 1 indicates a gasoline type, while 2 indicates a hybrid vehicle type. Notably, all hybrid vehicles considered in this study are categorized as hybrid electric vehicles (HEVs), which do not support external charging. Both the gasoline vehicles and HEVs utilize gasoline. Within these HEVs, batteries recharge through regenerative braking and the internal combustion engine, with the electric motor working in tandem with the gasoline engine. To conduct a reliable and generalizable analysis, numerous vehicle categories (26 combinations of model, vehicle name, and engine type) were taken into account, as shown in Table 1.

Table 1. Statistics of the vehicle categories (model, vehicle name, and engine type) and the corresponding numerical labels with the percentage share of each category. Engine Type 1 represents gasoline vehicles, while Engine Type 2 represents hybrid vehicles.

Category	Model	Vehicle Name	Engine Type	Count	Percentage (%)
1	1	8	1	16,101	7.05
2	1	9	1	3218	1.41
3	2	5	1	3938	1.73
4	3	2	2	3262	1.43
5	4	4	2	13,045	5.72
6	5	16	2	7279	3.19
7	6	3	2	5061	2.22
8	7	14	2	8562	3.75
9	8	18	1	6383	2.80
10	9	4	1	12,418	5.44
11	10	18	1	8532	3.74
12	11	19	1	6825	2.99
13	12	18	2	25,106	11.00
14	13	19	2	15,632	6.85
15	14	19	2	4525	1.98
16	15	7	2	9943	4.36
17	16	16	1	12,487	5.47
18	17	6	2	8918	3.91
19	18	13	1	3962	1.74
20	19	11	2	5388	2.36
21	20	17	2	21,196	9.29
22	21	12	2	4374	1.92
23	22	13	2	11,266	4.94
24	23	10	2	1804	0.79
25	23	15	2	1590	0.70
26	24	1	2	7466	3.27
Total				228,281	100.00

Table 2 illustrates the novel driving behavior data proposed in this study for V_{fuel} . The dataset includes the counts of dangerous behaviors of speeding, sudden accelerating, sudden braking, alongside driving distance [m] on the expressways, national highways, and local roads, respectively, on a monthly basis, in addition to the monthly total driving time [s] and monthly total driving distance [m] on all types of roads, which were recorded by the telematics on the vehicles. Moreover, the monthly fuel consumption [L] was recorded using the internal measurement devices of the vehicles. The monthly CO_2 emission E_{CO_2} was then calculated as a secondary parameter based on the assumption that engines emit 2.3 kg of CO_2 for every 1 L of gasoline consumed [29,30], and that this relationship holds true for all vehicles under consideration. The E_{CO_2} can be calculated using Equation (1). It is important to note that the E_{CO_2} values were not directly verified by measurements, but rather estimated based on the assumed linear relationship between fuel consumption and CO_2 emissions.

$$E_{\text{CO}_2} [\text{kg}] = V_{\text{fuel}} [\text{L}] \times 2.3 [\text{kg/L}] \quad (1)$$

The speeding is determined based on the average speed over a certain distance, with thresholds set according to the type of road (expressways, national highways, local roads). On general roads, the average speed is typically measured from one traffic signal to the next. On expressways and national highways, where there are no traffic signals, it is measured over distances of about 2000 m. The speeding thresholds are set at 120 km/h for expressways and 80 km/h for national highways and local roads, respectively. Brief increases in speed for overtaking are not considered as speeding. Speeding is only determined if the average speed over a specific section exceeds the set thresholds. Sudden accelerating

and braking are defined as increasing or decreasing speed by 10 km/h or more within 1 s. However, accelerating and braking that occur in normal driving scenarios are not considered as sudden accelerating or sudden braking.

Table 2. Novel driving behavior data utilized for E_{CO_2} in this study.

Driving Behavior Data (Monthly)
Total driving time [s]
Total driving distance [m]
Counts of dangerous speeding on expressways [times]
Counts of dangerous sudden accelerating on expressways [times]
Counts of dangerous sudden braking on expressways [times]
Driving distance on expressways [m]
Counts of dangerous speeding on national highways [times]
Counts of dangerous sudden accelerating on national highways [times]
Counts of dangerous sudden braking on national highways [times]
Driving distance on national highways [m]
Counts of dangerous speeding on local roads [times]
Counts of dangerous sudden accelerating on local roads [times]
Counts of dangerous sudden braking on local roads [times]
Driving distance on local roads [m]

2.2. Multiple Feature Sets

2.2.1. Estimation for All Vehicle Categories

To analyze and compare the efficacy of various features for E_{CO_2} estimation, multiple feature sets were defined as follows. First, the monthly total driving time and total driving distance were combined to form the Base features. Second, the driving distance on the expressways, national highways, and local roads, respectively, were combined to form the D features. Third, the dangerous driving behaviors, including the counts of speeding, sudden accelerating, and sudden braking on the expressways, national highways, and local roads, respectively, were combined to form the B features. Fourth, the specification of the vehicles, including the model, vehicle name, and engine type were combined to form the S features. Finally, 8 feature sets were defined in total by different combinations of the D, B, and S feature sets to the Base feature set: Base, BaseD, BaseB, BaseS, BaseDB, BaseDS, BaseBS, BaseDBS, as shown below.

- Base features (Base feature set) consisted of the total driving time and total driving distance.
- D features consisted of driving distance on expressways, national highways, and local roads.
- B features consisted of dangerous driving behavior data on expressways, national highways, and local roads.
- S features consisted of specification of the vehicles (model, vehicle name, and engine type).
- BaseD feature set consisted of Base and D features.
- BaseB feature set consisted of Base and B features.
- BaseS feature set consisted of Base and S features.
- BaseDB feature set consisted of Base, D, and B features.
- BaseDS feature set consisted of Base, D, and S features.
- BaseBS feature set consisted of Base, B, and S features.
- BaseDBS feature set consisted of Base, D, B, and S features.

2.2.2. Estimation for Each Vehicle Category

To verify the efficacy and generalizability of the E_{CO_2} estimation model that mixes data from all vehicle categories together, individual E_{CO_2} estimation models for each vehicle category (model, vehicle name, and engine type) were defined separately using all the driving behavior features (BaseDB) listed in Table 1.

2.3. Dimensionality Reduction for Data Visualization

In order to elucidate underlying patterns in the high-dimensional dataset and simultaneously gain insights into the significance of individual features, we employed dimensionality reduction techniques to the complete feature set (BaseDBS). This approach not only allows for clear visualization of the intrinsic structure of the data in a lower-dimensional space to highlight clusters, outliers, and trends, but also provides insights into feature contributions. By analyzing the principal components, as in principal component analysis (PCA), or embedding vectors in methods like t-Distributed Stochastic Neighbor Embedding (t-SNE) [31] or Uniform Manifold Approximation and Projection (UMAP) [32], we can discern which features are most influential in shaping the observed patterns, deepening our comprehension of data relationships. The PCA, a commonly used linear dimensionality reduction technique, offers valuable insights for data visualization. However, its linear nature limits its ability to capture non-linear data relationships. Conversely, t-SNE, another popular choice for dimensionality reduction and visualization, excels in handling non-linearities. Additionally, we employed UMAP, a versatile manifold learning method adept at non-linear dimensionality reduction, which is also more computationally efficient than t-SNE. While both t-SNE and UMAP are non-linear, preserving local or high-dimensional data structures, t-SNE's computational complexity stands at $O(n^2)$, making it suboptimal for larger datasets. UMAP, with its approximate complexity of $O(n)$, is considerably more efficient for extensive datasets. For our study, PCA, t-SNE, and UMAP were applied to the BaseDBS feature set to condense the data dimensions from 19 to 2. Implementations of PCA and t-SNE were sourced from the Python Scikit-learn package [33,34], while UMAP was executed using its dedicated Python package [35]. We retained default hyper-parameters for all the three dimensionality reduction algorithms (PCA: 'n_components = 2'; t-SNE: 'learning_rate = auto' and 'perplexity = 30.0'; UMAP: 'n_neighbors = 15', 'min_dist = 0.1', and 'n_components = 2').

3. Regression Models

An MLR model and a random forest regression (RFR) model were conducted to evaluate the validity of each feature set defined in Section 2.2 for estimating the E_{CO_2} . Prior to fitting the MLR and RFR models, the features were normalized to ensure that each feature contributes proportionately to the result and used as explanatory variables. Concurrently, the E_{CO_2} served as the target variable. We used 80% of the total data points as the training set and the remaining 20% as the testing set after shuffling. The MLR and RFR models were implemented using the Python Scikit-learn package. The MLR model assumes a linear relationship between the target and explanatory variables while the RFR model can account for non-linearities. A 5-fold cross-validation was conducted to the RFR model by utilizing a grid search technique using the training set data to select the optimal hyper-parameters of the RFR model, as illustrated in Figure 1. In the RFR model, the hyper-parameter candidates were selected empirically as follows: 'n_estimators': [200, 600, 1000]; 'max_features': ['auto', 'sqrt']; 'max_depth': [10, 20, 30]. To gauge the performance of the model, the coefficient of determination (R^2), root mean squared error (RMSE), and mean absolute error (MAE) were used as the performance metrics. The calculation equations of the R^2 , RMSE, and MAE are shown in Equations (2)–(4). In Equations (2)–(4), y represents the target variable of the regression, \hat{y} represents the predicted value of y , n stands for the total number of data points, and \bar{y} is the average value of y spanning the n data points.

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (2)$$

$$RMSE(y, \hat{y}) = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \quad (3)$$

$$MAE(y, \hat{y}) = \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{n} \quad (4)$$

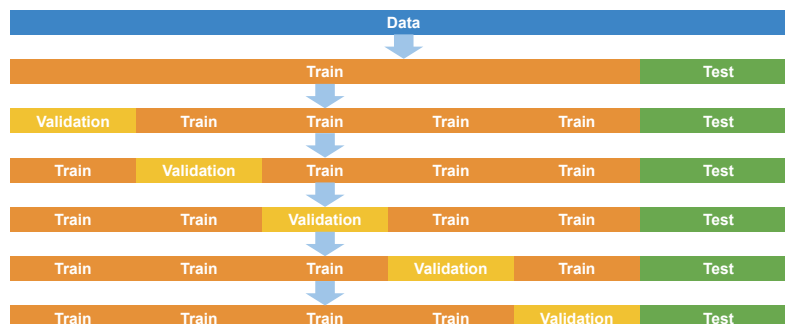


Figure 1. Training, validation, and testing sets' separation.

4. Results and Discussion

4.1. Data Visualization

Figure 2 depicts data visualizations derived from the PCA, t-SNE, and UMAP dimensionality reduction algorithms, utilizing the complete feature set (BaseDBS). The dimension-reduced data were color-coded using each feature separately for enhanced visualization, as shown in Figure 2. As can be observed in Figure 2, it is evident that the clusters in the PCA visualization are smaller than those in the t-SNE and UMAP visualizations, which makes it challenging to provide a detailed visualization. In Figure 2, UMAP gives a better visualization than t-SNE in terms of clarity. While t-SNE primarily aims to preserve the local structure of the data, sometimes at the expense of global structure, UMAP seeks to maintain a balance between local and global structures, often resulting in a more interpretable embedding. The distinction among different vehicle categories is markedly pronounced in the UMAP visualization presented in Figure 2, particularly when the data are color-coded by the model, vehicle name, and engine type. These categories exhibit distinct and clear borders in the embedding, indicating a significant influence of these features on the data structure. This clarity in demarcation underscores the critical role of the model, vehicle name, and engine type in differentiating driving characteristics across vehicle categories. Notably, this observation aligns with the findings from our RFR model, where the model, vehicle name, and engine type emerged as the top three features in terms of feature importance. This alignment reinforces the validity of our dimensionality reduction analysis and provides empirical support for the significant impact these features have on the ability of the RFR model to predict CO₂ emissions accurately. The precise delineation of clusters by these features in the UMAP visualization not only corroborates their importance but also highlights the efficacy of the UMAP in capturing the nuanced distinctions among vehicle categories, thereby offering valuable insights into the underlying structure of the data.

Figure 3 shows the boxplots of the driving behavior data and the E_{CO_2} of all data points, offering a visual representation of distribution and variability. Each boxplot delineates the interquartile range (IQR), highlighting the middle 50% of the data, with the lower and upper bounds of the box corresponding to the first and third quartiles, respectively. Inside the box, a horizontal line marks the median. The two bars above and below the box are the upper and lower whiskers, which extend from the first and third quartiles to the highest and lowest data points within 1.5 times the IQR, indicating the spread of the bulk of the data and highlighting outliers beyond this range. This visualization reveals a pattern of generally low-frequency dangerous driving behaviors among most users, punctuated by a small number of significant outliers. This observation prompts a deeper analysis of the outliers, suggesting potential areas for targeted interventions to mitigate high-risk driving behaviors. Furthermore, the clear demarcation between typical and outlier behaviors aids in understanding the relationship between driving habits and CO₂ emissions, underscoring the importance of addressing dangerous driving behaviors to enhance E_{CO_2} estimation.

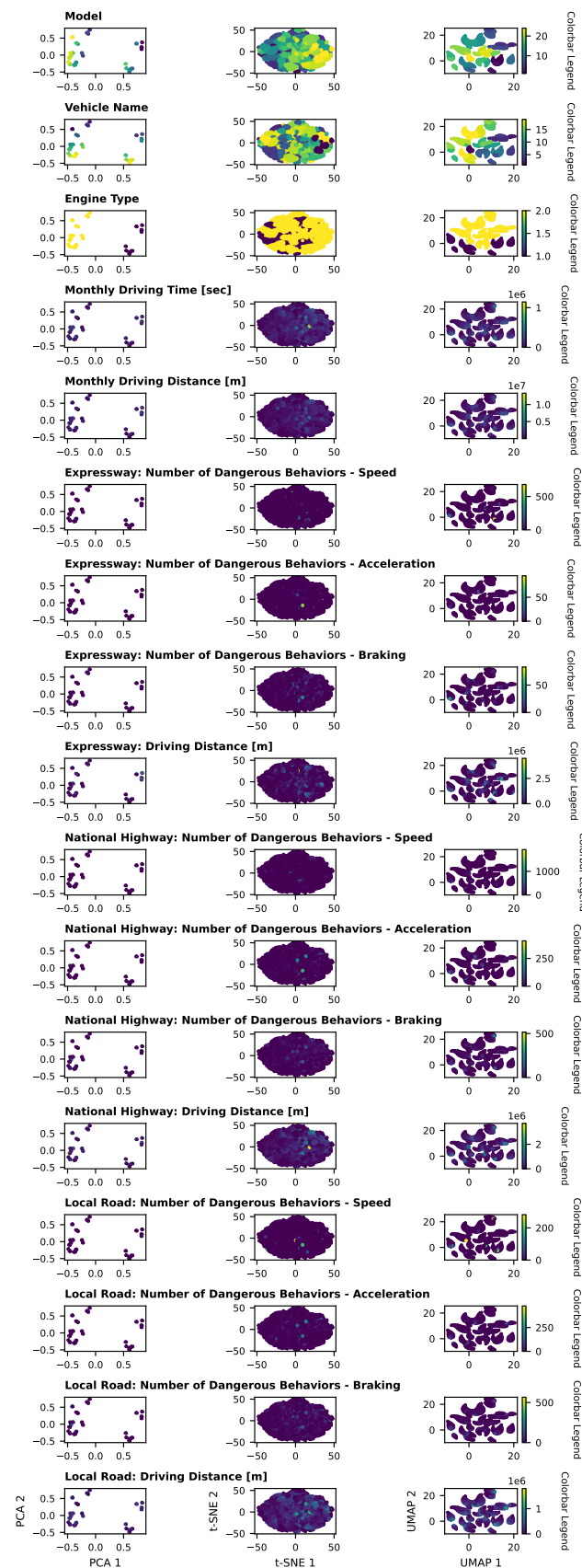


Figure 2. Data visualizations using PCA, t-SNE, and UMAP dimensionality reduction algorithms for all data points.

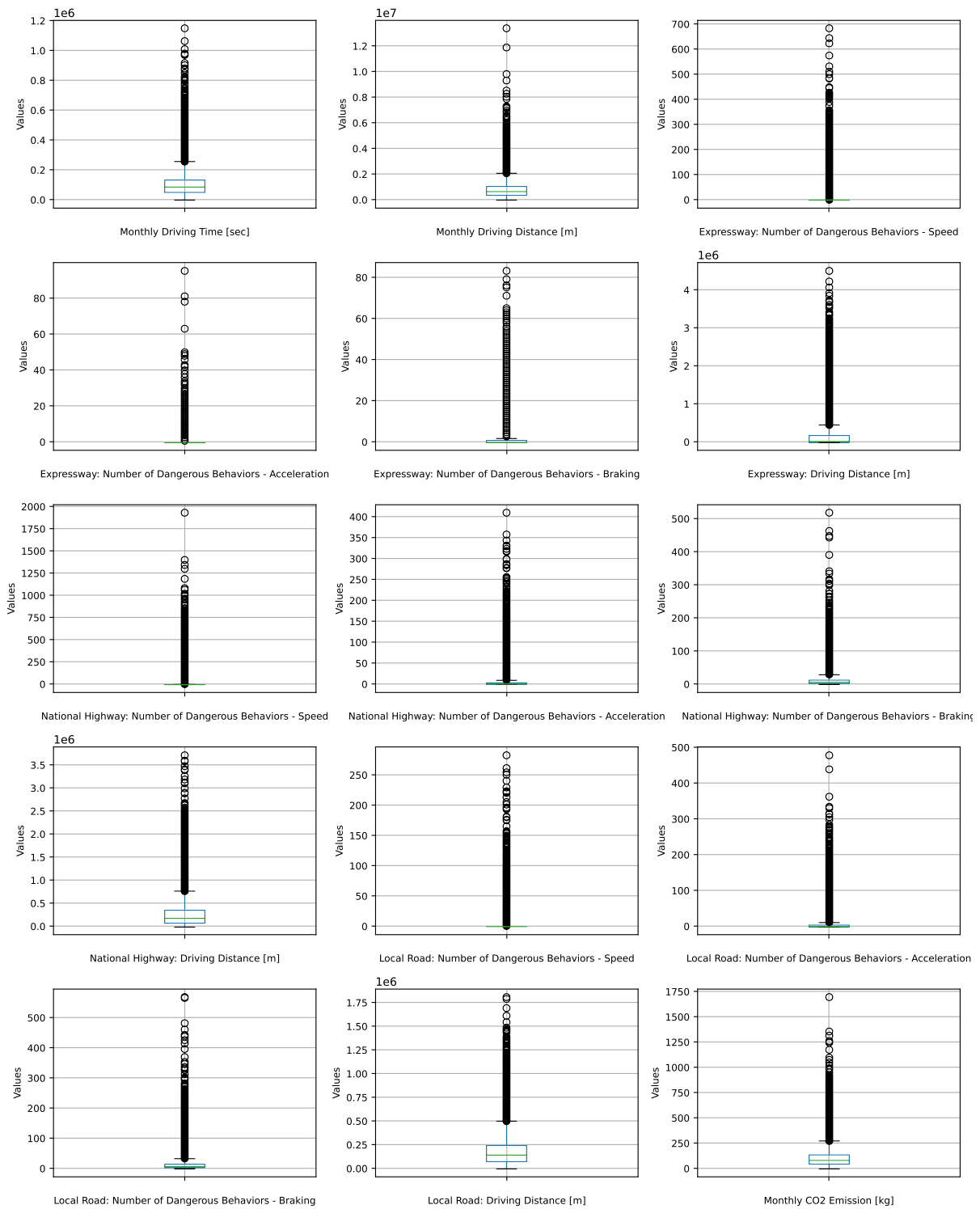


Figure 3. Boxplots of the driving behavior data and the E_{CO_2} for all data points.

4.2. Correlation Analysis

Figure 4 shows the heatmap of correlations among the driving behavior data and the E_{CO_2} . As is shown in Figure 4, the E_{CO_2} highly correlates with the total driving time and total driving distance, and moderately correlates with the driving distance on the three types of roads, while it weakly correlates with the dangerous driving behaviors on the three types of roads, implying the validity of the proposed novel driving behavior data for E_{CO_2} .

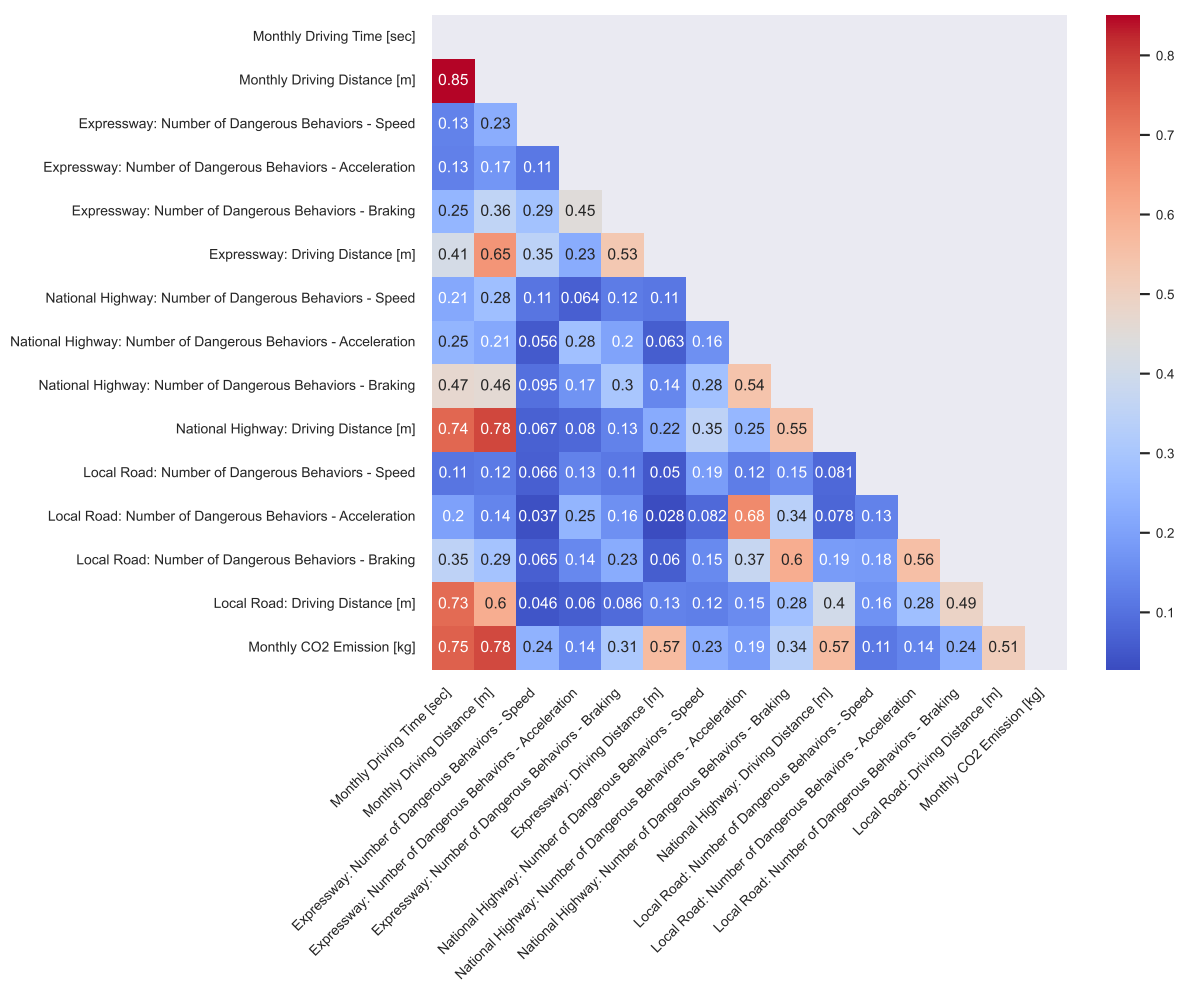


Figure 4. Correlation heatmap of the driving behavior data and the E_{CO_2} .

4.3. Machine Learning Regression Results

Table 3 presents the E_{CO_2} prediction performance metrics for each vehicle category separately. The performance metrics for estimating E_{CO_2} , using the eight feature sets defined in Section 2.2.1, are shown in Table 4. Figure 5 shows the observations-versus-predictions plots of E_{CO_2} predictions using the BaseDBS feature set by MLR and RFR models. As observed from Tables 3 and 4, the prediction accuracy using the BaseDBS feature set for all vehicle categories combined is even higher than the average prediction accuracy for each vehicle category when considered separately. This suggests the validity and feasibility of the method that combines all vehicle categories to create a generalizable model. The decision to evaluate all the eight feature sets on a generalized estimation model, rather than on individual vehicle models, is underpinned by the universally applicable relationship between driving behaviors and engine efficiency [20,21] and CO_2 emissions, as discussed in Section 1.

Observing Table 4, significant results can be summarized as follows:

1. The BaseD, BaseB, BaseS feature sets outperformed the Base feature set.
2. The BaseDB feature set outperformed both the BaseD and BaseB feature sets.
3. The BaseDS feature set outperformed both the BaseD and BaseS feature sets.
4. The BaseBS feature set outperformed both the BaseB and BaseS feature sets.
5. The BaseDBS feature set outperformed all the other feature sets.

Therefore, it has been confirmed that the proposed driving behavior features are relevant for E_{CO_2} modeling. Notably, the BaseDBS feature set yielded the highest prediction accuracy for both MLR and RFR models, with R^2 values reaching 0.842 and 0.975,

respectively. On the other hand, Figure 2 shows that, although no notable differences are observed when a single driving behavior feature is used for color-coding the embedding, combining these features together yields high performance for E_{CO_2} estimation, indicating the non-linearities of the data. Moreover, the RFR model outperformed the MLR model across all feature sets, as the RFR model can account for non-linearities, as indicated in Table 4 and Figure 5. The error distributions, illustrated in Figure 6, show a prominent peak centered around zero, indicating that most predictions closely match observed values. The errors appear to follow a normal distribution, with fewer large deviations, suggesting that the models are well-calibrated and reliable.

Table 3. The E_{CO_2} prediction performance of the MLR and RFR models using the BaseDB feature set for each vehicle category separately.

Category	MLR			RFR		
	R2	RMSE	MAE	R2	RMSE	MAE
1	0.967	23.394	15.977	0.975	20.378	14.255
2	0.961	24.458	15.875	0.957	25.751	17.176
3	0.976	17.357	11.315	0.973	18.448	11.826
4	0.973	11.233	7.838	0.972	11.466	7.665
5	0.946	18.996	12.069	0.963	15.696	10.162
6	0.970	12.238	8.297	0.963	13.526	8.725
7	0.946	16.323	7.842	0.935	17.922	8.191
8	0.972	11.808	7.821	0.969	12.387	7.485
9	0.967	10.342	6.800	0.967	10.426	6.671
10	0.967	18.246	12.771	0.969	17.629	11.880
11	0.948	12.364	6.423	0.967	9.815	6.259
12	0.947	16.269	7.706	0.954	15.126	7.800
13	0.924	12.658	6.803	0.962	8.949	5.443
14	0.943	12.370	7.078	0.961	10.252	6.799
15	0.946	13.702	9.927	0.941	14.318	9.660
16	0.943	10.689	6.835	0.951	9.995	6.560
17	0.971	15.696	9.266	0.979	13.278	8.673
18	0.978	8.598	5.670	0.977	8.967	5.638
19	0.980	12.789	9.036	0.979	13.067	8.911
20	0.915	17.889	8.531	0.964	11.642	8.001
21	0.940	13.100	8.406	0.965	10.055	6.582
22	0.944	13.121	9.734	0.947	12.825	8.308
23	0.954	12.760	8.532	0.958	12.202	8.235
24	0.935	19.747	14.499	0.912	22.972	15.189
25	0.909	20.356	14.581	0.906	20.678	14.283
26	0.962	11.472	8.166	0.954	12.658	8.254
Average	0.953	14.922	9.531	0.958	14.247	9.178

Table 4. The E_{CO_2} prediction performance of the MLR and RFR models using the proposed eight feature sets for all vehicle categories combined.

Feature Sets	MLR			RF		
	R2	RMSE	MAE	R2	RMSE	MAE
Base	0.632	50.933	33.875	0.642	50.232	33.452
BaseD	0.654	49.418	33.090	0.676	47.836	31.956
BaseB	0.647	49.879	33.396	0.686	47.034	31.439
BaseS	0.826	34.994	23.498	0.965	15.671	9.963
BaseDB	0.660	48.972	32.851	0.703	45.771	30.486
BaseDS	0.839	33.741	22.713	0.972	14.116	8.965
BaseBS	0.835	34.139	22.992	0.972	13.934	8.820
BaseDBS	0.842	33.387	22.456	0.975	13.293	8.329

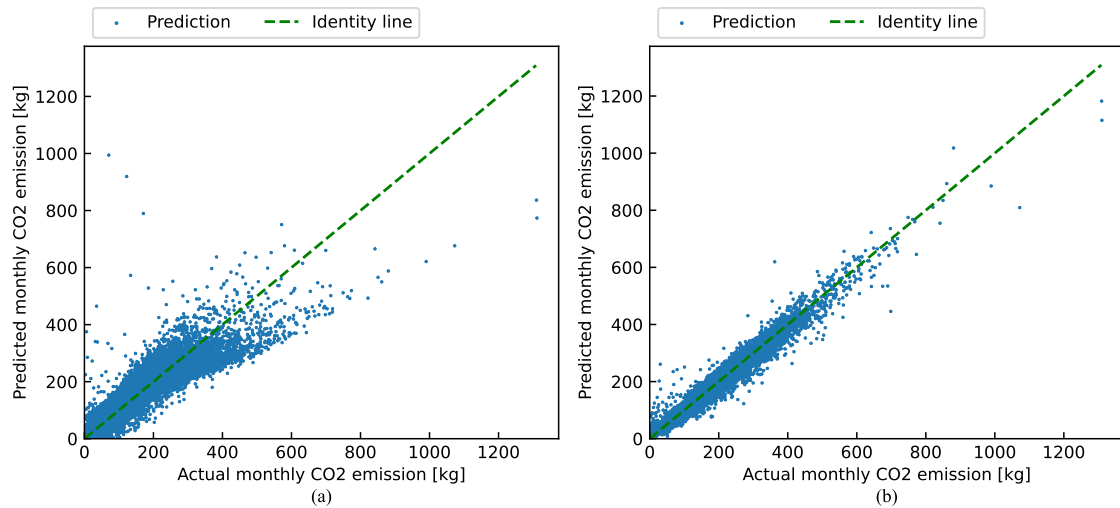


Figure 5. Observations versus predictions plots for E_{CO_2} prediction using the BaseDBS feature set by the MLR model (a) and the RFR model (b).

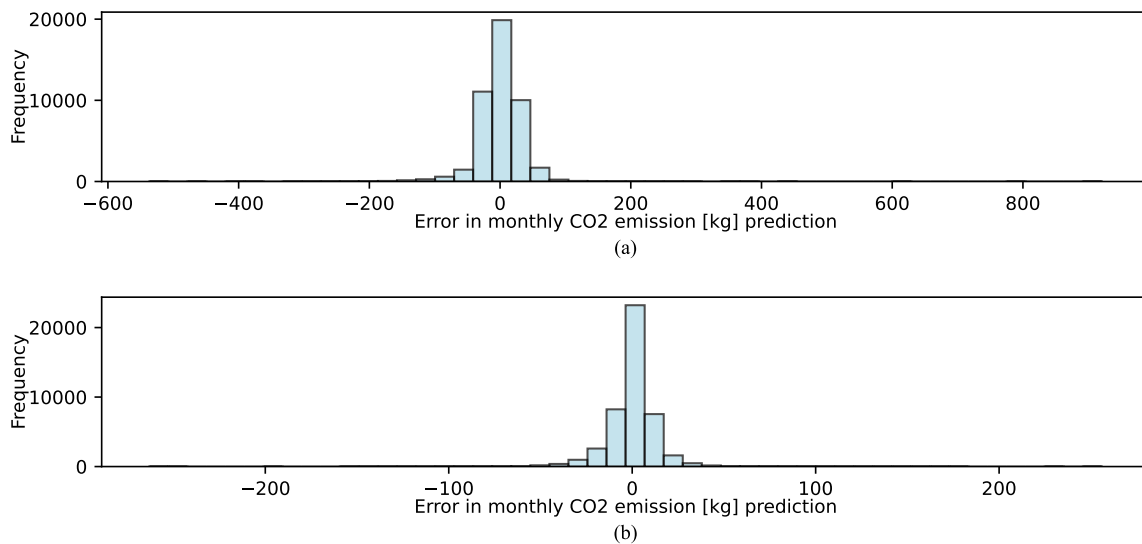


Figure 6. Error distributions for E_{CO_2} prediction using the BaseDBS feature set by the MLR model (a) and the RFR model (b).

Figure 7 displays the impurity-based feature importances of the explanatory variables for E_{CO_2} estimation using the BaseDBS feature set in the RFR regression model. As shown in Figure 7, the S features (model, vehicle name, and engine type), and Base features (total driving time and total driving distance) are ranked as the top five features in terms of feature importance. This aligns with the dimensionality reduction results mentioned in Section 4.1. Subsequently, the B features (driving behavior data on expressways, national highways, and local roads) are ranked 6th to 8th, 10th to 12th, and 14th to 16th, respectively. This ranking is considered reasonable, as according to general knowledge, vehicle speeds on expressways are typically higher than those on national highways, and speeds on national highways are generally higher than those on local roads. Higher speeds result in higher CO_2 emission rates, thereby having a greater impact on the feature importance ranking. Furthermore, the B features are ranked even higher than the D features (driving distances on expressways, national highways, and local roads), highlighting the validity of the proposed novel dangerous driving behavior features for E_{CO_2} estimation. This finding aligns with the RFR regression result, which showed that the BaseB feature set outperformed the BaseD feature set, as illustrated in Table 4. Therefore, the overall ranking is S features > Base features > B features > D features.

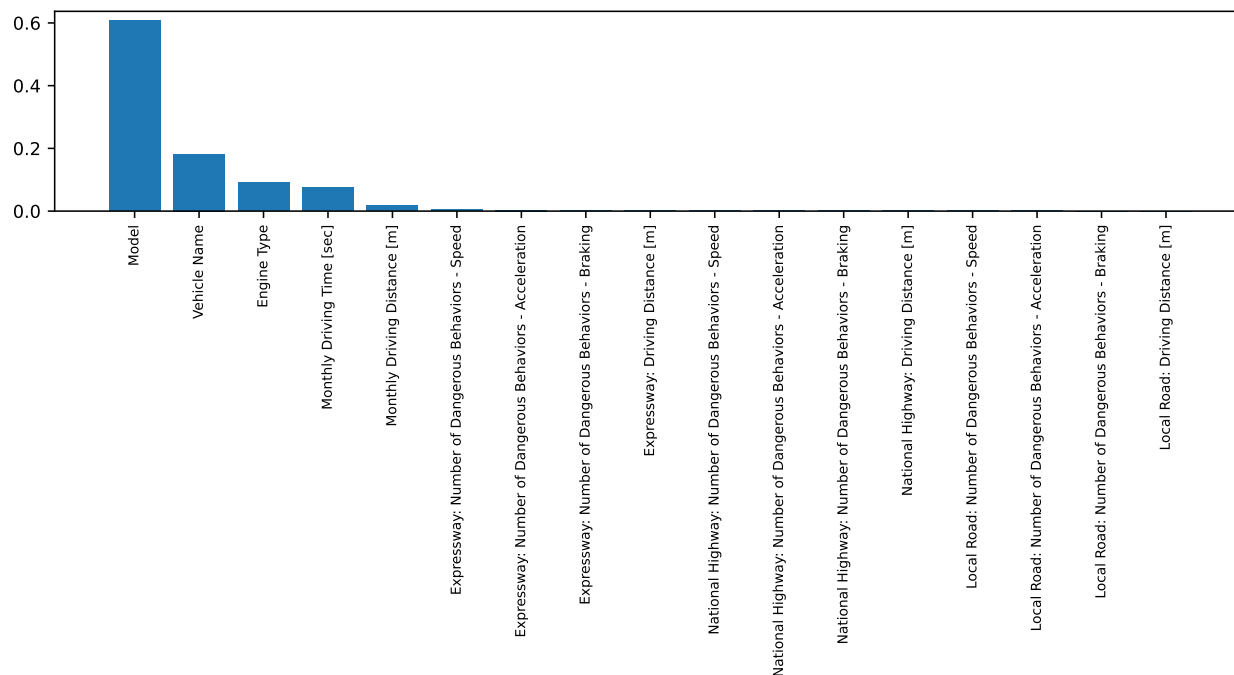


Figure 7. Feature importances for E_{CO_2} prediction using the BaseDBS feature set across all vehicle categories.

5. Conclusions and Future Work

This study has developed a novel approach for estimating monthly CO_2 emissions from individual gasoline vehicles and HEVs (both utilize gasoline) using regression algorithms, incorporating unprecedentedly detailed driving behavior data across diverse categories of TOYOTA cars. Our analysis revealed that the driving behavior data proposed in this study; specifically, the Base features (total driving time and driving distance), D features (driving distances on expressways, national highways, and local roads), B features (dangerous driving behavior data, including counts of sudden acceleration, sudden braking, and speeding on expressways, national highways, and local roads), and S features (specification of the vehicles, including model, vehicle name, and engine type) all have a positive influence on monthly CO_2 emission estimations. The overall ranking of the proposed features is as follows: S features > Base features > B features > D features. Utilizing a comprehensive feature set (BaseDBS), the RFR model achieved the highest prediction accuracy, with R^2 , RMSE, and MAE values of 0.975, 13.293 kg, and 8.329 kg, respectively, for predicting monthly CO_2 emissions. These results not only underscore the critical relevance of all examined driving behaviors to CO_2 emissions, but also highlight the superior predictive accuracy and generalizability across vehicle categories.

In future work, instead of encoding specific models and vehicle names into integer numerical labels (e.g., 1, 2, 3, ...) for inclusion in the regression models, it might be beneficial to incorporate general physical quantities such as the frontal projection area and the weight of the vehicles. This approach could enhance the generalizability of CO_2 emission estimation models to unknown models and vehicle names. Additionally, by integrating the CO_2 emission estimation model with incentive-based programs like the J-Credit Scheme or incorporating it into driver feedback systems, there is potential to significantly influence driving behaviors towards more energy-efficient practices, thereby contributing to broader environmental sustainability goals. It is important to note that the CO_2 emissions in this study were not directly verified by measurements, but rather estimated based on the assumed linear relationship with fuel consumption. Future work could involve directly measuring CO_2 emissions to validate and refine this relationship. Furthermore, consideration will also be extended to diesel vehicles, addressing the distinct characteristics and environmental impact of diesel fuel usage in addition to petrol vehicles,

to encompass a more comprehensive analysis of CO₂ emissions across different fuel types. In conclusion, the comprehensive approach to CO₂ emission estimation presented in this study offers a robust foundation for both advancing scientific understanding and developing practical solutions to the pressing challenge of vehicular emissions.

Author Contributions: Conceptualization, S.N.; methodology, Z.W.; software, Z.W.; validation, M.M.; formal analysis, Z.W.; investigation, Z.W. and M.M.; resources, S.N.; data curation, M.M.; writing—original draft preparation, Z.W.; writing—review and editing, Z.W. and M.M.; visualization, Z.W.; supervision, R.M.; project administration, R.M.; funding acquisition, S.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Aioi Nissay Dowa Insurance Co., Ltd.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: Author Shoma Nishimura was employed by the company Aioi Nissay Dowa Insurance Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. This work was sponsored in part by Aioi Nissay Dowa Insurance Co., Ltd. The funding sponsor provided the telematics data (driving behavior data) essential for our research and proposed a hypothesis on the correlation between individual driving behavior data and monthly individual vehicular CO₂ emissions. While the sponsor facilitated the collection of data and offered a hypothesis for investigation, they had no role in the design of the study beyond these contributions; in the analysis or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

IEA	International Energy Agency
EPA	The U.S. Environmental Protection Agency
CARB	California Air Resources Board
RPM	Revolutions per minute
HEV	Hybrid electric vehicle
Base	Total driving time and driving distance
D	Driving distances on expressways, national highways, and local roads
B	Dangerous driving behavior data on expressways, national highways, and local roads
S	Specification of the vehicles (model, vehicle name, and engine type)
BaseD	Base and D features
BaseB	Base and B features
BaseS	Base and S features
BaseDB	Base, D and B features
BaseDS	Base, D, and S features
BaseBS	Base, B, and S features
BaseDBS	Base, D, B, and S features
PCA	Principal component analysis
t-SNE	t-distributed Stochastic Neighbor Embedding
UMAP	Uniform Manifold Approximation and Projection
MLR	Multiple linear regression
RFR	Random forest regression
R ²	Coefficient of determination
RMSE	Root mean squared error
MAE	Mean absolute error
IQR	Interquartile range
D_{driving}	Monthly driving distance [m]
V_{fuel}	Monthly fuel consumption [L]
E_{CO_2}	Monthly CO ₂ emission [kg]
y	Target variable
\hat{y}	Predicted target variable
n	Sequence length of the target variable
\bar{y}	Average value of y spanning the n data points
i	i -th value in a variable sequence

References

1. European Commission. Road Transport: Reducing CO₂ Emissions from Vehicles. 2020. Available online: https://climate.ec.europa.eu/eu-action/transport-emissions/road-transport-reducing-co2-emissions-vehicles_en (accessed on 24 January 2024).
2. Wei, F.; Zhang, X.; Chu, J.; Yang, F.; Yuan, Z. Energy and environmental efficiency of China's transportation sectors considering CO₂ emission uncertainty. *Transp. Res. Part D Trans. Environ.* **2021**, *97*, 102955. [CrossRef]
3. Eco-Mo Foundation. 2020 Transport and Environment in Japan. 2020. Available online: <https://www.ecomo.or.jp/english/pdf/tej2020.pdf> (accessed on 21 January 2024).
4. International Energy Agency. Cars and Vans. Available online: <https://www.iea.org/energy-system/transport/cars-and-vans> (accessed on 29 January 2024).
5. Karczewski, M.; Chojnowski, J.; Szamrej, G. A Review of Low-CO₂ Emission Fuels for a Dual-Fuel RCCI Engine. *Energies* **2021**, *14*, 5067. [CrossRef]
6. Dziubak, T.; Karczewski, M. Experimental Studies of the Effect of Air Filter Pressure Drop on the Composition and Emission Changes of a Compression Ignition Internal Combustion Engine. *Energies* **2022**, *15*, 4815. [CrossRef]
7. Dziubak, T. Theoretical and Experimental Studies of Uneven Dust Suction from a Multi-Cyclone Settling Tank in a Two-Stage Air Filter. *Energies* **2021**, *14*, 8396. [CrossRef]
8. Dziubak, T.; Karczewski, M. Experimental Study of the Effect of Air Filter Pressure Drop on Internal Combustion Engine Performance. *Energies* **2022**, *15*, 3285. [CrossRef]
9. Tian, X.; Geng, Y.; Zhong, S.; Wilson, J.; Gao, C.; Chen, W.; Yu, Z.; Hao, H. A bibliometric analysis on trends and characters of carbon emissions from transport sector. *Transp. Res. Part D Trans. Environ.* **2018**, *59*, 1–10. [CrossRef]
10. Sperling, D.; Gordon, D. *Two Billion Cars: Driving toward Sustainability*; Oxford University Press: Oxford, UK, 2009.
11. Sterner, T. Fuel taxes: An important instrument for climate policy. *Energy Policy* **2007**, *35*, 3194–3202. [CrossRef]
12. Dargay, J.; Gately, D.; Sommer, M. Vehicle ownership and income growth, worldwide: 1960–2030. *Energy J.* **2007**, *28*, 143–170. [CrossRef]
13. Ministry of Economy, Trade and Industry. J-Credit Scheme. 2022. Available online: <https://japancredit.go.jp/english/> (accessed on 1 June 2023).
14. United Nations Framework Convention on Climate Change. The Paris Agreement. Available online: <https://unfccc.int/process-and-meetings/the-paris-agreement> (accessed on 16 December 2023).
15. Zhang, K.; Batterman, S. Near-road air pollutant concentrations of CO and PM_{2.5}: A comparison of MOBILE6. 2/CALINE4 and generalized additive models. *Atmos. Environ.* **2010**, *44*, 1740–1748. [CrossRef]
16. California. Air Resources Board. *EMFAC7F*; The Board: Sacramento, CA, USA, 1993.
17. Oduro, S.D.; Metia, S.; Duc, H.; Ha, Q.P. CO₂ vehicular emission statistical analysis with instantaneous speed and acceleration as predictor variables. In Proceedings of the 2013 International Conference on Control, Automation and Information Sciences (ICCAIS), Nha Trang, Vietnam, 25–28 November 2013; pp. 158–163.
18. Ahn, K.; Rakha, H.; Trani, A.; Van Aerde, M. Estimating vehicle fuel consumption and emissions based on instantaneous speed and acceleration levels. *Period. Polytech. Trans. Eng.* **2002**, *128*, 182–190. [CrossRef]
19. Maroju, R.; Nishimura, S.; Wang, Z.; Matsushashi, R. Estimating Vehicular Fuel Consumption and CO₂ Emissions by Machine Learning Using Only Speed and Acceleration. *J. Jpn. Soc.* **2023**, *44*, 30–38.
20. Wróblewski, P. An Innovative Approach to Data Analysis in The Field of Energy Consumption and Energy Conversion Efficiency in Vehicle Drive Systems—The Impact of Operational and Utility Factors. In Proceedings of the 37th International Business Information Management Association (IBIMA), Cordoba, Spain, 30–31 May 2021; pp. 1–2.
21. Saxena, S.; Phadke, A.; Gopal, A. Understanding the fuel savings potential from deploying hybrid cars in China. *Appl. Energy* **2014**, *113*, 1127–1133. [CrossRef]
22. Lois, D.; Wang, Y.; Boggio-Marzet, A.; Monzon, A. Multivariate analysis of fuel consumption related to eco-driving: Interaction of driving patterns and external factors. *Transp. Res. Part D Trans. Environ.* **2019**, *72*, 232–242. [CrossRef]
23. Jiménez, J.L.; Valido, J.; Molden, N. The drivers behind differences between official and actual vehicle efficiency and CO₂ emissions. *Transp. Res. Part D Trans. Environ.* **2019**, *67*, 628–641. [CrossRef]
24. Mane, A.; Djordjevic, B.; Ghosh, B. A data-driven framework for incentivising fuel-efficient driving behaviour in heavy-duty vehicles. *Transp. Res. Part D Trans. Environ.* **2021**, *95*, 102845. [CrossRef]
25. Grote, M.; Williams, I.; Preston, J.; Kemp, S. Including congestion effects in urban road traffic CO₂ emissions modelling: Do Local Government Authorities have the right options? *Transp. Res. Part D Trans. Environ.* **2016**, *43*, 95–106. [CrossRef]
26. Sharifi, F.; Birt, A.G.; Gu, C.; Shelton, J.; Farzaneh, R.; Zietsman, J.; Fraser, A.; Chester, M. Regional CO₂ impact assessment of road infrastructure improvements. *Transp. Res. Part D Trans. Environ.* **2021**, *90*, 102638. [CrossRef]
27. Samaras, C.; Tsokolis, D.; Toffolo, S.; Magra, G.; Ntziachristos, L.; Samaras, Z. Improving fuel consumption and CO₂ emissions calculations in urban areas by coupling a dynamic micro traffic model with an instantaneous emissions model. *Transp. Res. Part D Trans. Environ.* **2018**, *65*, 772–783. [CrossRef]
28. Zhang, L.; Zhu, Z.; Zhang, Z.; Song, G.; Zhai, Z.; Yu, L. An improved method for evaluating eco-driving behavior based-on speed-specific vehicle-specific power distributions. *Transp. Res. Part D Trans. Environ.* **2022**, *113*, 103476. [CrossRef]
29. Jahirul, M.I.; Masjuki, H.H.; Saidur, R.; Kalam, M.A.; Jayed, M.H.; Wazed, M.A. Comparative engine performance and emission analysis of CNG and gasoline in a retrofitted car engine. *Appl. Therm. Eng.* **2010**, *30*, 2219–2226. [CrossRef]

30. Vitliemov, P.; Kolev, N.; Marinov, M. Economic evaluation of the implementation of policy actions in the field of energy efficiency. *Int. J. Energy Econ. Policy* **2019**, *9*, 106–113. [[CrossRef](#)]
31. van der Maaten, L.; Hinton, G. Visualizing Data Using t-SNE. 2008. Available online: <https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf?fbcl> (accessed on 21 January 2024).
32. McInnes, L.; Healy, J.; Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv* **2018**, arXiv:1802.03426. <https://doi.org/10.48550/arXiv.1802.03426>.
33. Scikit-Learn Developers. Sklearn.Decomposition.PCA. Available online: <https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html> (accessed on 16 December 2023).
34. Scikit-Learn Developers. Sklearn.Manifold.TSNE. Available online: <https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html> (accessed on 17 January 2024).
35. McInnes, L.; Healy, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction—Umap 0.5 Documentation. 2018. Available online: <https://umap-learn.readthedocs.io/en/latest/index.html> (accessed on 29 January 2024).

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.