IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 28, 2021, accepted July 7, 2021, date of publication July 12, 2021, date of current version July 16, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3096279

Analysis of Carbon Dioxide Emissions From Road Transport Using Taxi Trips

MOHAMMADHOSSEIN GHAHRAMANI, (Member, IEEE), AND FRANCESCO PILLA
Spatial Dynamics Laboratory, University College Dublin, Dublin 4, D04 V1W8 Ireland

Corresponding author: Mohammadhossein Ghahramani (sepehr.ghahramani@ucd.ie)

This work was supported in part by a Grant from the Science Foundation Ireland under Grant 16/IA/4610.

ABSTRACT Transport emissions, including road, rail, air, and marine transportation, account for a large part of the overall emissions; hence, there is a need to review strategies for managing associated issues and coping with negative impacts. A simultaneous improvement in economic efficiency can help us achieve our desired objectives in the concerned context. Sharing economy, i.e., a peer-to-peer-based sharing of access to assets, can help reduce the total resources required and consequently reduce carbon footprints. In line with this objective, we propose an intelligent model to study carbon dioxide emissions from road transport using taxi trips in Dublin, Ireland. The proposed method is a hybrid unsupervised learning approach tailored for the particular structure of the problem. We present how an intelligent approach can be implemented to model CO2 emissions from road transport. The model categorizes taxis based on different features related to the emissions they release. Five clusters are detected, which can be attributed to varying levels of emissions. Accordingly, those vehicles labeled as the highest emitters can be targeted for further improvements in reducing CO2, i.e., replacing pollutant cars with electric cars or including them in the taxi fleet as sharing ones only.

INDEX TERMS Artificial intelligence, CO2 reduction, energy consumption, sharing economy.

I. INTRODUCTION

Global CO2 emissions from fossil fuels have significantly increased in the last decades, causing different concerns, i.e., global warming, health impacts, and climate change, which implies economic and environmental complexity. Carbon emissions have increased by about 90% since 1970, with emissions from fossil fuel combustion and industrial processes contributing about 78% of the total greenhouse gas emissions increase. The continuous growth of economic activities leads to increased energy consumption and is accountable for the increasing in climate change related impacts and environmental degradation. The continuous increment in consumption of non-renewable resources will result in an increase in carbon dioxide emissions, with a further aggravated negative impact on the surrounding environment. While most countries are committed to limiting global warming and reducing greenhouse gas emissions, significant concerns should be addressed, and a question that comes into mind is, how can the world achieve this goal?

The associate editor coordinating the review of this manuscript and approving it for publication was Maurice J. Khabbaz.

Without transformation in how energy is used and immediate implementation of emissions cuts, limiting the progression of global warming will be difficult. Transport is one of the largest sources of energy demand and contributors to carbon emissions. This work focuses on transport emissions in Ireland, where traffic is responsible for more than 40% of Ireland's CO2 emissions. Within the sector, road transport is the largest emitter of GHG emissions. Greenhouse gas emissions come from different sources, i.e., Carbon Dioxide (CO2) and Methane. However, the most significant GHG emissions source (i.e., approximately 60% of all GHG emissions) in Ireland is CO2 emissions. Hence, there is an urgent need to explore causes and possible solutions to slow the growth of human-made emissions of carbon dioxide. Reducing emissions from this sector can be accomplished by increased public transport use, incentives for renewing the vehicle fleet, and shared mobility [1]. A considerable amount of CO2 emissions can be reduced by better managing fuel consumption. In doing so, new technologies such as Artificial Intelligence (AI), Big Data, Machine Learning, and Internet of Things (IoT) can improve policy makers' understanding of the causes of environmental problems [2]. The great strength of AI and algorithmic learning lies in their ability to take vast amounts

of seemingly unconnected data from the environment, intuiting connections that humans overlook, drawing insights, and recommending appropriate actions. Such approaches can be used to evaluate the impact of transportation policies on fuel consumption and CO2 emissions and assess the potential for their reduction.

It has been discussed that policies adopted in the past decades have reduced fuel consumption and GHG emissions from the transportation sector. However, despite the immense magnitude and adaptability of transportation policies, they will not put the transportation sector on a trajectory to reduce emissions sufficiently to keep global average temperature increase below the agreed 2° C limit. It seems that the possibility of such reductions would require the adoption of a broader set of sustainable practices, along with a transformation of vehicle technologies and transportation systems [3]–[7]. In order to provide policy-makers with information on policy options and alternatives, several questions should be answered, i.e., what is the current growth rate in CO2 emissions from the transport sector by mode and region? What is the potential to reduce the carbon footprint further from the transportation sector? And how countries compare to one another in terms of vehicle efficiency and mode shares. Various research methodologies should be employed to answer these questions, among which quantitative and qualitative data analysis using computational intelligence methods.

As discussed, one potential solution might be a reduction in emissions relative to petrol and diesel vehicles. To achieve this goal, we need to investigate the emissions and reduction potential. We aim to use data analytics's mentioned capabilities and implement an optimized unsupervised learning model to expedite better decision-making. We propose an optimized approach to explore the underlying determinants of emissions change and study scenarios to estimate the potential of emission reduction. The insights can provide a deeper level of understanding regarding fuel consumption and the policy guidance for policymakers. As discussed, CO2 emissions induced by taxis account for a high proportion of air pollution. The availability of taxi data presents new opportunities for addressing CO2 emissions caused by taxis. Since the amount of CO2 a taxi emits is directly related to the amount of fuel it consumes, different features (e.g., engine displacement, city MPG, and highway MPG) related to each vehicle are considered in the proposed model. We implement an optimized solution to detect different car clusters in the taxi fleet and study their underlying pattern, given their characteristics. In this way, the most pollutant cars can be detected and replaced to maximize the impact on emissions if limited financial resources are available. The contributions of this work are as follows:

- The potential to reduce CO2 emissions in a collaborative consumption context from the transport sector in Dublin city is studied.
- An optimized data-driven model based on an unsupervised learning approach is tailored for assessing the carbon footprint.

The paper is organized as follows: some related work is presented in Section II; the proposed approach with its associated discussions is presented in Section III; Section IV shows the experimental results; and Section V concludes the paper.

II. RELATED WORK

CO2 emissions are a major contributor to climate change impacts; therefore, different research studies have been conducted to analyze and mitigate threats and tackle related concerns. In [8], the authors have studied the historical CO2 emissions of Bangladesh's electricity sector from 1979 to 2018. A logarithmic mean divisia index method has been implemented, given three distinct scenarios for predicting future emissions. The achieved results suggest that CO2 emissions will peak at 58.97 Mtoe by 2040 in the concerned country. Bamisile *et al.* have presented a detailed analysis of the causes, trends, and solutions to carbon emission in Africa [9]. First, they have investigated the impact of economic development on Africa's CO2 emissions trend. Then, an AI-based method based on a neural network has been developed to predict the future trend of total CO2 emission in the continent. Finally, renewable energy sources for power generation are explored as a viable solution for CO2 emission reduction in Africa. Spatiotemporal approaches [10]–[14] can be utilized to explore CO2 emissions. Several studies have analyzed emissions based on spatial decomposition methods [15], [16]. In [15] the spatiotemporal evolution of decoupling and driving factors of CO2 emissions of several countries is investigated based on an integrated model for the period 1991 to 2016. The results show that the decoupling statuses of higher-income countries are generally better than lower-income countries. Different spatiotemporal evolution of the driving factors of CO2 emissions is analyzed based on a Kaya identity and LMDI model. Besides the industrial sector, agricultural practices have been pointed out as key contributors to GHG emissions. A regression technique with non-additive fixed effects together with a quintile decomposition technique has been used to explore whether and to what extent the relationship between agricultural and economic factors differs across low, intermediate, and high CO2 emitters in [17]. The results reveal that the shift from traditional farming methods to mechanized farming has increased the amount of CO2 emitted.

Many studies have focused on the mitigation potential of CO2 emissions from transport sector [18]–[22]. Most of these studies are concerned with traditional decoupling analysis of CO2 emissions from the transport sector. In [23], the authors use Tapio and Log-Mean Divisia Index models based on an extended Kaya identity to study the decoupling relationship between energy-related carbon emissions and economic growth in Cameroon's transport sector. The same approach is used in [24] to illustrate the relationship between the development of the transport sector and its CO2 emissions in several provinces in China. Georgatzi *et al.* have employed an integrated model based on Fully-Modified OLS

TABLE 1. Some observations of the taxi trip dataset.

id_trip	id_driver	route_distance (m)	car_model	car_manufacturer	origin_lat	origin_lon	destination_lat	destination_lon	seats_count
152022208	20013484	1150	308	Peugeot	53.255795	-6.113086	53.247959	-6.124337	4
152018237	20247263	39212	Caddy	Volkswagen	53.346373	-6.282119	53.186332	-6.805687	6
152022088	31443338	2424	Prius	Toyota	53.409018	-6.399085	53.393114	-6.319501	4
152021942	23567304	5581	Fluence	Renault	53.330554	-6.377290	53.338521	-6.240921	4
152021875	26561331	7808	Avensis	Toyota	53.341149	-6.226603	53.354913	-6.244909	4

TABLE 2. Some observations of the fuel economy dataset.

Make	Model	Engine Displacement	City MPG	Highway MPG	Combined MPG	CO2 Emission (Grams/Mile)
Honda	Accord	1.8	22	27	24	370.31
Toyota	Prius	1.5	40	38	40	216.756
Volkswagen	Golf	1.8	31	40	34	289.411
Volkswagen	Jetta	1.8	19	26	22	403.95
Volvo	V60	2	25	36	29	302.13

and Dynamic OLS approaches to examine the relationship between CO2 emissions caused by the transport sector activity and their statistically significant determinants [25]. They have highlighted the importance of policies and technological innovation in the transition to a low carbon economy.

Recently, smart cities have become an interesting research topic for academics, industry, and government [26]–[30]. However, there is relatively limited research on using new technologies such as big data, AI, and data-driven techniques for microscopic vehicle emission modelling. Hence, in this work, we focus on applying such methods in carbon emissions management and control systems to explore the emission reduction potential in the transport sector.

III. UNSUPERVISED LEARNING MODEL

The initial step for implementing a data-driven model is to perform data cleansing and preprocessing. Since the quality of data affects the analysis, it is of the utmost importance to enhance the data quality via preprocessing operations [31]. When a dataset is obtaiend from different sources—like the dataset used in this work—its integration presents some particular challenges like styles of record-keeping and aggregation. Without handling these challenges, the analysis's result cannot be precise. Therefore, certain steps must be executed to convert a raw dataset into a clean one. Missing values, noises, and outliers should be identified and properly handled. The primary method for detecting outliers is to quantify, through some metrics, the extent to which a single data item deviates from the others in the dataset. We used outlier labelling methods, and all observations beyond predefined intervals were removed.

A. DATASETS

In this work, two datasets were used, i.e., data of taxi trip records obtained from an Irish taxi company and a fuel economy dataset. The taxi trip data contains millions of anonymous origin-destination records and different variables, i.e., driver's ID, distance travelled, car manufacturer, car model, and seat count. Some sample records are presented in Table 1. The fuel economy dataset also includes various features such as car manufacturer, car model, engine displacement, city MPG, highway MPG, combined MPG, and CO2 emission. Some sample records are presented in Table 2. Both datasets were aggregated and integrated based on their common features, i.e., car model and manufacturer. It is worth mentioning that taxi trip records did not include any details about cars' model year, but such information was available in the other dataset. Hence, features such as the mean value of city MPG, highway MPG, combined MPG, and CO2 emission for each car model were calculated. Moreover, some car models/manufacturers in the taxi trip data were missing in the fuel economy dataset. To deal with this concern, we performed a left-merge operation, and missing values were deleted. Such missing values accounted for less than 15% of the whole dataset. Given the "seat_count" variable, the number of seats ranges from two to eight seats. A new feature (i.e., Vehicle Class) was created and populated as "Small Cars", "Midsize Cars" and "Large Cars" according to the "seat_count" feature. Redundant features such as coordinates variables also were deleted. Such variables can be used for trajectory analysis which is out of the scope of this work. Since the travelled distance (i.e., route_distance variable) in each trip has been measured in meters, we converted it to miles. Then, the distance value was multiplied by the corresponding CO2 emission associated with each car. In this way, CO2 emission in each trip was measured. Each record of the integrated dataset can be defined as an m-tuple (m is the number of features).

Let matrix $X \in \mathbf{R}^{n\times m}$ as:

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \quad (1)$$

where $\mathbf{R}$ is the real number set, X_i is the i^{th} observation and its corresponding variables (m-tuple), and n is the number of records. The data were normalized such that each data point lies between a given minimum and maximum value. It should be mentioned that, due to the unsupervised nature of this work, we assumed that the variables have a Gaussian probability distribution, and power transform methods like Box-Cox were used to normalize our dataset. Note that all values are positive. We defined a Logarithmic transformation, which can be applied to non-normal distribution. Such transformation procedure aims to find some value for κ such that the transformed data is as close to normally distributed as possible.

$$x_i^{(\kappa)} = \begin{cases} \dfrac{x_i^{\kappa} - 1}{\kappa}, & \text{if } \kappa \geq 0 \\ \ln(x_i), & \text{if } \kappa = 0 \end{cases} \tag{2}$$

After the data is normalized, the unsupervised learning approach can be performed. Relevant discussions are presented next.

B. CLUSTERING APPROACH

We deal with an unsupervised learning problem in this work since the concerned data does not include any target feature. Different approaches can be implemented to model CO2 emissions from the road transport sector. Unsupervised learning approaches can be divided into two main categories: non-hierarchical and hierarchical methods. The former, i.e., partition-based and density-based techniques, aim at finding similar objects based on predefined parameters. They relocate objects among different clusters until a convergence criterion is satisfied. Density-based methods are often computationally expensive, and partition-based methods have approximate linear time complexity [32]. Such approaches may not be the best for our problem due to the need for initial configuration. Moreover, comparing the obtained results using a non-hierarchical method with others in the same category is challenging since each non-hierarchical technique has a specific attitude toward solving an unsupervised problem, and determining the clustering tendency is not trivial [33]. Hence, in this work, the latter was considered. We have proposed an optimized hierarchical algorithm and tailored it for analyzing CO2 emissions data from road transport. Different validity measures were tested to make sure the results are reliable. We have also compared the results obtained from the implemented algorithm with several hierarchical methods based on the Cophenetic correlation coefficient.

As discussed throughout the pre-processing section, the concerned data include different features. We aim to divide this dataset into relatively homogeneous clusters based on similarity measures among observations (trips). Generally speaking, hierarchical algorithms are categorized into divisive and agglomerative methods. These methods operate based on defining an inter-cluster distance among data points. They seek to conduct a hierarchy of clusters by splitting a dataset into different subsets [34]. The proposed algorithm treats each observation as a separate cluster in the initial phase of the learning process. These clusters are merged (based on different similarity/agglomeration measures) until only one cluster remains. A validity measure is then performed to find a proper number of clusters. It should be mentioned that most agglomeration measures are based on Euclidean and Manhattan distance. However, we have integrated a specific ranking procedure based on a weighted correlation mechanism to achieve more robust results. It is worth noting that different correlation coefficient metrics (i.e., Pearson, Kendall, and Spearman) have been also tested. All these metrics are used to measure the strength of association between two observations. However, we found that the weighted coefficient integrated into our model is more effective than the mentioned ones. Details regarding the ranking schema are explained next.

C. RANKING SCHEMA

Given the set of n observations consists of d features, i.e., trips and associated variables, a weighted rank-order correlation coefficient (ρ) can be defined to measure similarities among pairwise data points. For a set of observations (with n rows and d features), an $n \times d$ ranking matrix is defined:

$$R = \begin{bmatrix} r_{11} & r_{12} & r_{13} & \dots & r_{1d} \\ r_{21} & r_{22} & r_{23} & \dots & r_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & r_{n3} & \dots & r_{nd} \end{bmatrix} \tag{3}$$

where r_{ij} denotes the i^{th} rank of j^{th} feature. The ranking matrix consists of unique ranks from 1 to n. The correlations (ρ) are the permutation of pairwise observations given their rank orders, each of which represents the distances among all pairwise data points. It can be measured according to the following formulas:

$$\rho = \frac{\frac{1}{n}\sum_{i=1}^{n}(R_i - \bar{R}_i).(R_j - \bar{R}_j)}{\sqrt{(\frac{1}{n}\sum_{i=1}^{n}(R_i - \bar{R}_i)^2).(\frac{1}{n}\sum_{i=1}^{n}(R_i - \bar{R}_i)^2)}} \tag{4}$$

where R_i and R_j are the i^{th} and j^{th} rank-vectors, respectively. The ranks can be measured and stored in matrix R. Accordingly, $D = R_i - R_j$ in (4) denotes the difference between the ranks associated with the variables of the i^{th} and j^{th} observations. The correlation coefficient between R_i and R_j is given by

$$\rho_{R_iR_j} = 1 - \frac{2\sum_{i=1}^{n} D^2}{max(\sum_{i=1}^{n} D^2)} \tag{5}$$

where $max(\sum_{i=1}^{n} D^2) = \frac{n^3-n}{3}$. The normalized version of 5 can be defined

$$\rho_{R_iR_j} = 1 - \frac{2\sum_{k=1}^{n}\sum_{i=1}^{k}(R_i - R_j)}{max(\sum_{k=1}^{n}\sum_{i=1}^{k} D^2)(R_i - R_j)} \tag{6}$$

where $\sum_{k=1}^{n}\sum_{i=1}^{k}(R_i - R_j) = \frac{1}{2}\sum_{k=1}^{n} D^2$. Let $\phi_k = \sum_{i=1}^{k}(R_i - R_j)$, a weighted correlation coefficient can

Algorithm 1 The Pseudo-Code for the Unsupervised Model

Input: $X \leftarrow$ data observations;
$p \leftarrow |X|$ i.e., the number of observations;
$\omega \leftarrow []$ i.e., weights matrix; $\lambda \leftarrow []$ i.e., labels;
$l \leftarrow []$ i.e., position on the dendrogram;
$\Theta \leftarrow []$, i.e., Relationship data structure including ranks and weights;
Output: *Clusters*
function Rank(R_i, R_j)
  $R \leftarrow 1 + \frac{2\sum_{i=1}^{n}(R_i - R_j)(n+1-R_i)^p}{\sum_{i=1}^{n}(n+1-2R_i)(n+1-R_i)^p}$
  return R
end function
for $i \leftarrow 1$ to $n-1$ **do**
  $(X_\alpha, X_\beta) \leftarrow \arg\max \rho(R_i R_j)$
  $l_* \leftarrow (l_{\alpha\beta}, X_\alpha, \overset{R}{X_\beta})$
  $l \leftarrow Append(l_*)$
  $X_{new} \leftarrow X \backslash \{(X_\alpha, X_\beta\}$
  For all $X \in X_{new}$
  $\rho(l_*, X) = \rho(X, l_*) = InterClusterDist(X, l_*)$
  $l \leftarrow l \cup l_*$
end for
function InterClusterDist(*labels*, D^*)
  $Dist \leftarrow [\,]$
  $l \leftarrow [\,]$
  for $i \leftarrow 1$ to $n-1$ **do**
    $l_i \leftarrow l_i \backslash l_0$
    $Dist \leftarrow \arg\min_{l_0 \in l} D_i^*[l]$
    $l \leftarrow Append(Dist)$
  end for
end function
return *Clusters*

be defined

$$\omega = 1 - \frac{2\sum_{i=1}^{n} v_i \phi_i}{max(\sum_{i=1}^{n} v_i \phi_i)} \tag{7}$$

$$\omega_p^{R_i R_j} = 1 + \frac{2\sum_{i=1}^{n}(R_i - R_j)(n+1-R_i)^p}{\sum_{i=1}^{n}(n+1-2R_i)(n+1-R_i)^p} \tag{8}$$

As explained, the proposed unsupervised learning approach starts by dividing the taxi trips dataset into singleton clusters. Most similar observations were identified given the conducted method, and clusters were merged based on rank orders and the defined coefficient. The pseudo-code for the procedure is presented in Algorithm 1. It should be mentioned, $\omega_p^{R_i R_j}$ take values in the interval of $[-1, 1]$. The two extreme values reveal minimum distance (completely identical orderings) or maximum distance (opposite rankings). It yields 0 if there is no correlation between the rank orders. The inter-cluster dissimilarity measure was implemented to determine similarities among the created clusters. Generally speaking, different techniques can be used to define an inter-cluster dissimilarity measure, such as single, complete, average, and ward linkage. These agglomeration functions

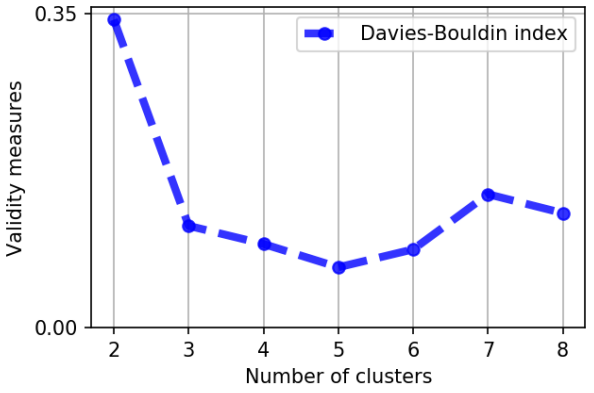

FIGURE 1. Validity measure to find an optimal number of clusters.

TABLE 3. Comparing the results of several hierarchical clustering methods given different settings.

Method	Linkage	Number of clusters	Cophenetic
Divisive	Single	6	0.723
Divisive	Complete	5	0.792
Divisive	Average	6	0.671
Divisive	Ward	5	0.665
Agglomerative	Single	5	0.813
Agglomerative	Complete	5	0.837
Agglomerative	Average	6	0.804
Agglomerative	Ward	5	0.839
Proposed method	**Number of clusters: 5**		**Cophenetic: 0.891**

are used to measure the similarity of newly formed clusters in hierarchical learning processes. Single-linkage function is based on the shortest distance among clusters while the largest distance is used in the Complete-linkage method (i.e., $d_c(C_I, C_J) = \max_{i \in C_I, j \in C_J} d_{ij}$, where C_I, C_J are two clusters). The average distance is measured and used to define similarity in the average-linkage agglomeration. We have integrated a weighted linkage function, and all different linkage methods were also tested. The obtained results are presented and compared next.

IV. RESULTS

All observations (taxi trips including distance traveled in each trip and fuel economy variables) were fed into the proposed unsupervised model as its input. Different filtering and normalization methods were implemented. A correlation coefficient was integrated into the model to measure the proximity of observations. A weighted-linkage method was employed on pairwise clusters. All the discussed procedures helped us obtain the hierarchy of clusters. However, the relationships among clusters have to be analyzed. Selecting an optimal number of clusters is a challenging task, and the quality of clustering depends on the optimal choice. Therefore, the Davies-Bouldin index was taken into account to adjust an appropriate cluster level and select a proper cutoff threshold. This index is based on the intra-cluster and inter-cluster variations. Let $C = (C_1, C_2, \ldots, C_k)$ be the identified

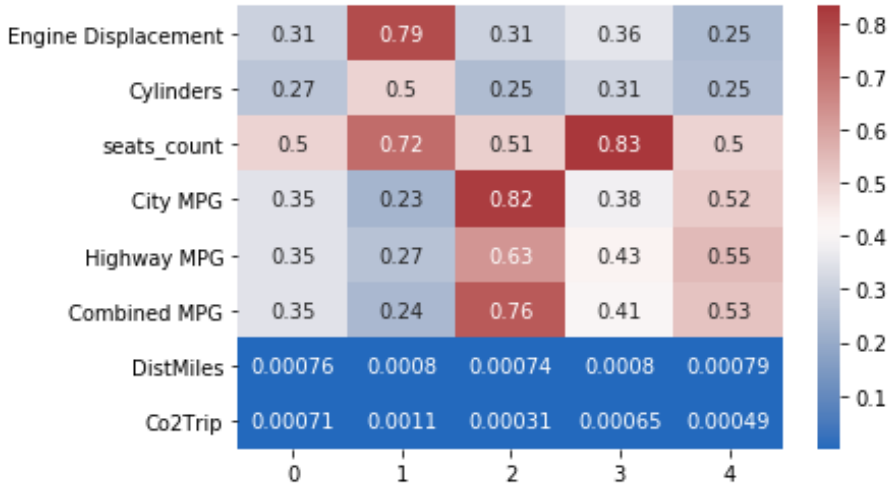

FIGURE 2. Comparisons of different features.

TABLE 4. Characteristics of different features in 5 detected clusters.

Cluster	No. Trips	Engine Displacement	Cylinders	seats_count	City MPG	Highway MPG	Combined MPG	DistMiles	Co2Trip
0	5100620	0.307623	0.273732	0.500360	0.348948	0.349357	0.350359	0.000763	0.000711
1	963254	0.786645	0.500254	0.720955	0.228552	0.265800	0.240257	0.000801	0.001095
2	2225495	0.308962	0.248937	0.508244	0.819402	0.629137	0.758614	0.000736	0.000311
3	1117487	0.355939	0.313153	0.832083	0.382698	0.430485	0.406645	0.000804	0.000645
4	1098780	0.245241	0.254799	0.500498	0.516122	0.551351	0.531621	0.000786	0.000491

k clusters and $\omega(R_i, R_j)$ be the defined correlation value between two observations, R_i and R_j. Let $C_p = \{C_1^p, C_2^p, \ldots, C_n^p\}$ be the p^{th} cluster including n observations, the Davies-Bouldin index, then, can be measured as follows:

$$D^*(k) = \frac{1}{k}\sum_{i=1}^{k} max(\frac{\Omega(C_i) + \Omega(C_j)}{\overline{D}(C_i, C_j)}) \tag{9}$$

where $\overline{D}(C_i, C_j)$ is the intra-cluster distance between C_i and C_j clusters. $\Omega(C_i)$ and $\Omega(C_j)$ are the inter-cluster distance among observations in C_i and C_j clusters respectively and are defined as

$$\Omega(C_i) = \frac{\sum_{\Phi_z \in C_i} \omega(\Phi_z, c_i)}{|C_i|}, \quad i = 1, 2, \ldots, p \tag{10}$$

where Φ is a centriod of each clusters.

Fig. 9 illustrates the results achieved from the DBI metric. According to the results, we choosed 5 as the optimal number of clusters. We also implemented and tested various hierarchical unsupervised learning methods (i.e., Divisive and Agglomerative clustering), including different linkage criteria. The DBI metric has been employed to choose an optimal number of clusters for all the tested methods. The results are illustrated in Table 3.

The Cophenetic coefficient has been employed in order to compare the results achieved from the proposed model with others tested in this work. The metric calculates the average inter-cluster distances between pairwise observations compared to their actual distances.

$$\frac{\sum_{i \neq j}(\overline{D}(x_i, x_j) - \hat{D})}{\sqrt{\sum_{i \neq j}[\overline{D}(x_i, x_j) - \hat{D})^2] \sum_{i \neq j}[\overline{C}(x_i, x_j) - \hat{C})^2]}} \tag{11}$$

where D represents average distances among all pairwise clusters and C represents the Cophenetic distance.

The calculated values for each method are presented in Table 3. As can be seen, the calculated measure for the proposed model is 0.891, slightly higher than those tested methods. Given the results, we can make sure the proposed model is reliable. Since the model is perfectly fitted to the purpose, we can safely run it to obtain clusters and identify underlying patterns associated with carbon emission footprints generated. According to the optimal number of clusters (i.e., 5 in this work), the clustering results obtained from the taxi trip dataset are presented in Fig. 2 and Table 4. Note that the given values have been normalized (as explained throughout the paper). The number of trips in each cluster is also presented in Table 4. Given the results, Cluster 1 and the corresponding trips are the most contributors to carbon emissions. The vehicles in this cluster have relatively large engines. Interestingly, the average number of seats in

Cluster 3 is larger than in other clusters, but the average carbon emission footprint is low. We can conclude that vehicles in Cluster 1 can be identified and replaced with electric cars to reduce carbon emissions in the concerned taxi fleet.

V. CONCLUSION

Recent studies have indicated that there are correlations between economic growth and environmental sustainability. The sharing economy concept, i.e., shared use of service, can be a valuable solution to environmental concerns. It can have positive impacts by reducing overall energy inputs and helps reduce pollutants, emissions, and carbon footprints. The concept, i.e., collaborative consumption, focuses on sharing underutilized assets to improve efficiency and sustainability. Reducing resource consumption is a significant contribution of the sharing economy towards the sustainability agenda.

The transport sector is an integral part of the total economic system and a large consumer of energy. It contributes significantly to Ireland's greenhouse gas emissions. The emissions induced by taxis account for a high proportion. In this paper, we proposed an unsupervised learning approach to investigate the impact of taxi trips on CO2 emissions. An optimized hierarchical clustering model was implemented to identify clusters associated with emissions. In this way, the most polluting trips were identified. The vehicles associated with these trips (i.e., the highest priority carbon emissions) can be identified to take future decisions. Identifying the cars related to these trips can provide authorities with several possible paths of action, or alternatives, i.e., replacing them with electric vehicles. The model can be used to define effective policies such as incentive mechanisms for electric vehicles. It helps understand how to automatically infer patterns associated with CO2 emissions in a collaborative consumption context. Governments have limited amounts of resources. Hence, they can target specific clusters of vehicles to be replaced with electric vehicles and maximize the impact on carbon emissions with limited financial resources. Such analyses help to adapt their strategies according to climate scenarios and have a clear understanding of the potential ways to reduce consumption-based emissions.

ACKNOWLEDGMENT

For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

REFERENCES

[1] F. Bistaffa, C. Blum, J. Cerquides, A. Farinelli, and J. A. Rodriguez-Aguilar, "A computational approach to quantify the benefits of ridesharing for policy makers and travellers," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 1, pp. 119–130, Jan. 2021, doi: 10.1109/TITS.2019.2954982.

[2] M. Ghahramani, A. O'Hagan, M. Zhou, and J. Sweeney, "Intelligent geodemographic clustering based on neural network and particle swarm optimization," *IEEE Trans. Syst., Man, Cybern. Syst.*, early access, May 3, 2021, doi: 10.1109/TSMC.2021.3072357.

[3] Y. Liu and J. Tan, "Green traffic-oriented heavy-duty vehicle emission characteristics of China VI based on portable emission measurement systems," *IEEE Access*, vol. 8, pp. 106639–106647, 2020, doi: 10.1109/ACCESS.2020.3000665.

[4] Q. Zhang, F. Li, F. Long, and Q. Ling, "Vehicle emission forecasting based on wavelet transform and long short-term memory network," *IEEE Access*, vol. 6, pp. 56984–56994, 2018, doi: 10.1109/ACCESS.2018.2874068.

[5] F. Chen, Z. Yin, Y. Ye, and D. Sun, "Taxi hailing choice behavior and economic benefit analysis of emission reduction based on multi-mode travel big data," *Transp. Policy*, vol. 97, pp. 73–84, Oct. 2020, doi: 10.1016/j.tranpol.2020.04.001.

[6] D. J. Sun and X. Ding, "Spatiotemporal evolution of ridesourcing markets under the new restriction policy: A case study in Shanghai," *Transp. Res. A, Pract. Police*, vol. 130, pp. 227–239, Dec. 2019, doi: 10.1016/j.tra.2019.09.052.

[7] D. J. Sun, K. Zhang, and S. Shen, "Analyzing spatiotemporal traffic line source emissions based on massive Didi online car-hailing service data," *Transp. Res. D, Transp. Environ.*, vol. 62, pp. 699–714, Jul. 2018, doi: 10.1016/j.trd.2018.04.024.

[8] M. M. Hasan and W. Chongbo, "Estimating energy-related CO_2 emission growth in Bangladesh: The LMDI decomposition method approach," *Energy Strategy Rev.*, vol. 32, Nov. 2020, Art. no. 100565, doi: 10.1016/j.esr.2020.100565.

[9] O. Bamisile, S. Obiora, Q. Huang, N. Yimen, I. A. Idriss, D. Cai, and M. Dagbasi, "Impact of economic development on CO_2 emission in Africa; the role of BEVs and hydrogen production in renewable energy integration," *Int. J. Hydrogen Energy*, vol. 46, no. 2, pp. 2755–2773, Jan. 2021, doi: 10.1016/j.ijhydene.2020.10.134.

[10] P. D'Orey and M. Ferreira, "ITS for sustainable mobility: A survey on applications and impact assessment tools," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 2, pp. 477–493, Apr. 2014, doi: 10.1109/TITS.2013.2287257.

[11] S. Hu, A. O'Hagan, J. Sweeney, and M. Ghahramani, "A spatial machine learning model for analysing customers' lapse behaviour in life insurance," *Ann. Actuarial Sci.*, vol. 15, pp. 367–393, Nov. 2020, doi: 10.1017/S1748499520000329.

[12] F. Pilati, I. Zennaro, D. Battini, and A. Persona, "The sustainable parcel delivery (SPD) problem: Economic and environmental considerations for 3PLs," *IEEE Access*, vol. 8, pp. 71880–71892, 2020, doi: 10.1109/ACCESS.2020.2987380.

[13] M. Ghahramani, M. Zhou, and C. T. Hon, "Extracting significant mobile phone interaction patterns based on community structures," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 3, pp. 1031–1041, Mar. 2019, doi: 10.1109/TITS.2018.2836800.

[14] M. Ghahramani, M. Zhou, and C. T. Hon, "Mobile phone data analysis: A spatial exploration toward hotspot detection," *IEEE Trans. Autom. Sci. Eng.*, vol. 16, no. 1, pp. 351–362, Jan. 2019, doi: 10.1109/TASE.2018.2795241.

[15] M. Hu, R. Li, W. You, Y. Liu, and C.-C. Lee, "Spatiotemporal evolution of decoupling and driving forces of CO_2 emissions on economic growth along the belt and road," *J. Cleaner Prod.*, vol. 277, Dec. 2020, Art. no. 123272, doi: 10.1016/j.jclepro.2020.123272.

[16] X. Yu, Z. Liang, J. Fan, J. Zhang, Y. Luo, and X. Zhu, "Spatial decomposition of city-level CO_2 emission changes in Beijing-Tianjin-Hebei," *J. Cleaner Prod.*, vol. 296, May 2021, Art. no. 126613, doi: 10.1016/j.jclepro.2021.126613.

[17] I. D. Nwaka, M. U. Nwogu, K. E. Uma, and G. N. Ike, "Agricultural production and CO_2 emissions from two sources in the ECOWAS region: New insights from quantile regression and decomposition analysis," *Sci. Total Environ.*, vol. 748, Dec. 2020, Art. no. 141329, doi: 10.1016/j.scitotenv.2020.141329.

[18] Y. Song, M. Zhang, and C. Shan, "Research on the decoupling trend and mitigation potential of CO_2 emissions from China's transport sector," *Energy*, vol. 183, pp. 837–843, Sep. 2019, doi: 10.1016/j.energy.2019.07.011.

[19] D. Nicolaides, D. Cebon, and J. Miles, "An urban charging infrastructure for electric road freight operations: A case study for cambridge UK," *IEEE Syst. J.*, vol. 13, no. 2, pp. 2057–2068, Jun. 2019, doi: 10.1109/JSYST.2018.2864693.

[20] M. Pourakbari-Kasmaei, M. Lehtonen, J. Contreras, and J. R. S. Mantovani, "Carbon footprint management: A pathway toward smart emission abatement," *IEEE Trans. Ind. Informat.*, vol. 16, no. 2, pp. 935–948, Feb. 2020, doi: 10.1109/TII.2019.2922394.

[21] Q. Hou, Y. Guan, and S. Yu, "Stochastic differential game model analysis of emission-reduction technology under cost-sharing contracts in the carbon trading market," *IEEE Access*, vol. 8, pp. 167328–167340, 2020, doi: 10.1109/ACCESS.2020.3023391.

[22] Q. Wang, Z. Huang, J. Hu, and Z. Liu, "A carbon emission evaluation method for remanufacturing process of a used vehicle CVT gearbox," *IEEE Access*, vol. 8, pp. 193257–193267, 2020, doi: 10.1109/ACCESS.2020.3027709.
[23] J. Engo, "Decoupling analysis of CO_2 emissions from transport sector in cameroon," *Sustain. Cities Soc.*, vol. 51, Nov. 2019, Art. no. 101732, doi: 10.1016/j.scs.2019.101732.
[24] Y. Li, Q. Du, X. Lu, J. Wu, and X. Han, "Relationship between the development and CO_2 emissions of transport sector in China," *Transp. Res. D, Transp. Environ.*, vol. 74, pp. 1–14, Sep. 2019, doi: 10.1016/j.trd.2019.07.011.
[25] V. V. Georgatzi, Y. Stamboulis, and A. Vetsikas, "Examining the determinants of CO_2 emissions caused by the transport sector: Empirical evidence from 12 European countries," *Econ. Anal. Policy*, vol. 65, pp. 11–20, Mar. 2020, doi: 10.1016/j.eap.2019.11.003.
[26] J. Yang, Y. Kwon, and D. Kim, "Regional smart city development focus: The South Korean national strategic smart city program," *IEEE Access*, vol. 9, pp. 7193–7210, 2021, doi: 10.1109/ACCESS.2020.3047139.
[27] M. Ghahramani, N. J. Galle, F. Duarte, C. Ratti, and F. Pilla, "Leveraging artificial intelligence to analyze citizens' opinions on urban green space," *City Environ. Interact.*, vol. 10, Apr. 2021, Art. no. 100058, doi: 10.1016/j.cacint.2021.100058.
[28] M. Ghahramani and F. Pilla, "Leveraging artificial intelligence to analyze the COVID-19 distribution pattern based on socio-economic determinants," *Sustain. Cities Soc.*, vol. 69, Jun. 2021, Art. no. 102848, doi: 10.1016/j.scs.2021.102848.
[29] M. Ghahramani, M. Zhou, and G. Wang, "Urban sensing based on mobile phone data: Approaches, applications, and challenges," *IEEE/CAA J. Automatica Sinica*, vol. 7, no. 3, pp. 627–637, May 2020, doi: 10.1109/JAS.2020.1003120.
[30] A. Kirimtat, O. Krejcar, A. Kertesz, and M. F. Tasgetiren, "Future trends and current state of smart city concepts: A survey," *IEEE Access*, vol. 8, pp. 86448–86467, 2020, doi: 10.1109/ACCESS.2020.2992441.
[31] M. Ghahramani, Y. Qiao, M. Zhou, A. O. Hagan, and J. Sweeney, "AI-based modeling and data-driven evaluation for smart manufacturing processes," *IEEE/CAA J. Automatica Sinica*, vol. 7, no. 4, pp. 1026–1037, Jul. 2020, doi: 10.1109/JAS.2020.1003114.
[32] C.-M. Huang, Y.-C. Huang, S.-J. Chen, and S.-P. Yang, "A hierarchical optimization method for parameter estimation of diesel generators," *IEEE Access*, vol. 8, pp. 176467–176479, 2020, doi: 10.1109/ACCESS.2020.3026670.
[33] H. Khaleghzadeh, R. R. Manumachu, and A. Lastovetsky, "A hierarchical data-partitioning algorithm for performance optimization of data-parallel applications on heterogeneous multi-accelerator NUMA nodes," *IEEE Access*, vol. 8, pp. 7861–7876, 2020, doi: 10.1109/ACCESS.2019.2959905.
[34] M. Takizawa and M. Yukawa, "Joint learning of model parameters and coefficients for online nonlinear estimation," *IEEE Access*, vol. 9, pp. 24026–24040, 2021, doi: 10.1109/ACCESS.2021.3053651.

MOHAMMADHOSSEIN GHAHRAMANI (Member, IEEE) received the M.S. degree in information technology engineering from the Amirkabir University of Technology (Tehran Polytechnic), Iran, and the Ph.D. degree in computer technology and application from the Macau University of Science and Technology, Macau, in 2018. From 2008 to 2014, he was a Technical Manager and a Senior Data Analyst of the Information Center of Institute for Research in Fundamental Sciences. He was a member of the Insight Centre for Data Analytics, University College Dublin (UCD), Ireland. He is currently a Research Fellow with UCD. He has over ten peer-reviewed journal articles as a first author. His research interests include smart cities, smart manufacturing, machine learning, artificial intelligence, big data, and the IoT. He has served the community by reviewing more than 100 articles for top journals in the last four years and has been active in organizing international conferences. He is the Co-Chair of the IEEE SMCS Technical Committee on AI-based Smart Manufacturing Systems.

FRANCESCO PILLA is currently an Associate Professor of smart cities with UCD, Ireland. His work lies at the intersection between cities and technologies with the goal to build better cities through technology, innovation, and citizen participation. His area of expertise is smart cities and in specific geospatial analysis and modeling of urban dynamics, which involves the development of GIS-based models and decision support tools, in order to pre-empt the impacts resulting from the interactions between human population and the environment. He uses a range of pervasive and community sensing applications as a means of calibration and validation of these GIS tools. He then integrates the GIS models and data streams from pervasive sensing deployments with advanced machine learning algorithms to gain a better understanding of the spatial dynamics in cities.