J Inf Process Syst, Vol.13, No.2, pp.285~304, April 2017
https://doi.org/10.3745/JIPS.04.0031

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

DTG Big Data Analysis for Fuel Consumption Estimation

Wonhee Cho* and Eunmi Choi*,**

Abstract

Big data information and pattern analysis have applications in many industrial sectors. To reduce energy consumption effectively, the eco-driving method that reduces the fuel consumption of vehicles has recently come under scrutiny. Using big data on commercial vehicles obtained from digital tachographs (DTGs), it is possible not only to aid traffic safety but also improve eco-driving. In this study, we estimate fuel consumption efficiency by processing and analyzing DTG big data for commercial vehicles using parallel processing with the MapReduce mechanism. Compared to the conventional measurement of fuel consumption using the On-Board Diagnostics II (OBD-II) device, in this paper, we use actual DTG data and OBD-II fuel consumption data to identify meaningful relationships to calculate fuel efficiency rates. Based on the driving pattern extracted from DTG data, estimating fuel consumption is possible by analyzing driving patterns obtained only from DTG big data.

Keywords

Big Data Analysis, DTG, Eco-Driving, Fuel Economy, Fuel Consumption Estimation, MapReduce

1. Introduction

Big data analytic methodology has been used to derive important predictions or estimations in various areas. The reduction of greenhouse gas emissions and energy consumption in the automotive sector has emerged as a global issue [1]. Eco-driving methods that can be used to save energy by improving automobile driving behavior represent major solutions. Eco-driving helps drivers improve their driving behavior and save fuel by monitoring fuel consumption in real time [2]. Accurate data about fuel consumption is required for fuel economy calculations. Thus, it is necessary to mount terminals such as On-Board Diagnostics version II (OBD-II) to calculate fuel consumption accurately, which is, however, expensive. As an alternative, big data from digital tachographs (DTGs) can be used as it is now mandatory to mount DTG on all commercial vehicles to reduce traffic accidents in Korea [3]. Vast vehicle-driving data regarding speed, revolutions per minute (RPM), GPS coordinates, acceleration, and brake signal are collected through DTG devices [4]. However, because DTG data do not contain fuel consumption information, studies requiring fuel consumption calculations could not be performed.

In this study, we attempt to build fuel consumption calculation formulae using only DTG data. Here,

Manuscript received July 14, 2015; first revision November 24, 2016; accepted December 4, 2016.

Corresponding Author: Eunmi Choi (emchoi@kookmin.ac.kr)
* Graduate School of Business IT, Kookmin University, Seoul, Korea (danylight@gmail.com)
** School of Software, College of Computer Science, Kookmin University, Seoul, Korea (emchoi@kookmin.ac.kr)

compared to the conventional measurement of fuel consumption using the OBD-II device, we propose a method for measuring fuel efficiency rates using only the driving pattern extracted from DTG data. We collect fuel consumption data during 5 months of normal vehicle running using 8 commercial vehicles mounted with OBD-II devices. We derive the driving pattern, basic statistical data, and information on dangerous driving behavior from the obtained DTG big data. We set the fuel consumption variable as the target and the driving pattern as independent, and calculate the formula using regression analysis. We evaluate this formula and compare the calculated fuel consumption values with the estimated values. Thus, it becomes possible to estimate fuel consumption using only DTG data and thereby guide drivers toward economic operation of their vehicles.

Using actual DTG data and the OBD-II fuel consumption formula, we identify the meaningful relationships between fuel consumption and the obtained DTG data that allow estimating the fuel efficiency rates. We apply parallel processing with the MapReduce mechanism for big data processing of the generated driving patterns. Based on the driving pattern extracted from the DTG data, it is possible to evaluate the fuel consumption estimation by analyzing driving patterns out of the DTG big data.

This paper is composed of the following five sections. In Section 2, we review the recent eco-driving research and fuel consumption estimation methods. In Section 3, we introduce DTG data properties, and review DTG big data analysis methods. In Section 4, we present fuel consumption effects of DTG data under big data analysis environments. We study micro-analysis methods using fuel consumption and DTG raw data. In addition, we perform basic data analysis of the driving pattern. In Section 5, we derive fuel economy formula and validate the model. In Section 6, we summarize the work of this paper and the direction of future work.

2. Eco-Driving Review

2.1 Eco-Driving Trend Review

Recently, the depletion of fossil fuels and global warming due to increasing carbon dioxide (CO_2) emissions have drawn attention around the auto industry, which has come under increasing scrutiny [5]. Thus, several efforts have been made toward reducing automobiles' fuel consumption and CO_2 emissions, according to the imposed regulations. In particular, the fuel efficiency of commercial vehicles such as cargo trucks is low; hence, considerable research effort has been put to addressing this issue. Moreover, the overall improvement of cars, such as through the development of hybrid cars, aims toward the prevention of global warming and the reduction of energy consumption.

On the other hand, eco-driving has also been investigated in order to induce improvements in driving behavior. Eco-driving is a cost-effective and quick method for improving fuel economy; in addition, eco-driving is a policy evaluated by international organizations such as the International Energy Agency (IEA) with the highest-priority of policy instruments of the Ministry of Transport inquiry for restriction on the use of fossil fuels [6]. By monitoring the driving habits of the driver and presenting the driver with a score to assess how often they engage in dangerous driving habits such as high average speed, sudden starts and stops, sudden acceleration, and sudden deceleration, research on ways of improving these driving habits in order to reduce excessive fuel consumption can be furthered [7].

2.2 The Eco-Driving Concept and Review

In a narrow sense, eco-driving shows only improved driving methods to the driver [6]. In a broad sense, the eco-driving concept includes green transportation, which is the mechanical improvement of the car, making changes to the poor driving skill of the driver, and promoting the use of green transportation methods such as the bicycle. It is determined that using eco-driving worldwide would lead to a reduction in CO_2 emissions of about 10% and a potential reduction in effective fuel consumption by 10%–20%. In the short term, the results of eco-driving education by country indicated a fuel consumption reduction of 5%–15% after training. In the most outstanding case, a fuel consumption improvement of 20%–50% could be achieved. Cars have four driving modes: starting, running, deceleration, and parking. In general, they consume 34% of the fuel when starting, 44% when driving, 7% when decelerating, and 15% when stopping. During normal driving, 1 km driving consumes 98.9 mL of fuel; 33.9 mL is consumed at the start, and it is possible to save 24.2 mL on departure while eco-driving. Moreover, a slow start reduces the total fuel consumption by 9.7%. In the driving stage, when changes in speed are increased, fuel consumption is considerably larger. It is possible to reduce fuel consumption by maintaining a constant speed while driving on a highway, because increasing the speed by 10 km/h on the highway can consume approximately 10% extra fuel. Boriboonsomsin et al. [1] measured fuel savings of about 6% in city streets and 1% in highways, when information related to fuel spending and driving manners is provided to the driver via an on-board eco-driving device.

2.3 Review of Fuel Consumption Estimation in Accordance with the Pattern of Driving (Fuel Efficiency)

Fuel consumption calculations from an OBD-II interface device, if there is such a device capable of directly measuring fuel consumption, allows accurate measurement. If the vehicle does not have a device with an OBD-II interface, it is difficult to estimate fuel consumption using only DTG data. However, there has been a lot of research over a long time on estimating fuel consumption by analyzing driving trajectory.

Regarding fuel consumption rates, there are multiple ways of analysis: using the driving speed and acceleration, adding vehicle or road information (slope and road status), or using the driver's driving pattern. The methods—VT-Micro [8], VT-Meso [9], Won et al. [10], and the Korea Transportation Safety Authority (KTSA) [11]—used speed information. CMEM [12] used both speed and vehicle information. VSP and MOVES [13] combine road slope information with speed information. Son et al. [14], Ericsson [15], Kang et al. [16] used driving pattern information.

Compared to other research, in this paper, we used speed, RPM, and the driving pattern obtained from the DTG. Before showing our new approach, we introduce the other recent research work as follows.

2.3.1 Study of the relationship between driving speed and fuel consumption

The fuel consumption estimation method uses vehicle speed [10], and it was developed by the Korea Automobile Testing & Research Institute of TS (KTSA) [11]. The following formulae show the fuel consumption estimation scheme for different vehicle types, as introduced to transportation facility investment evaluation guidelines.

Passenger Car	$: Lc = -0.00325338V^2 + 0.47782761V + 2.28593762$
Small Bus	$: L_{sb} = -0.00250760V^2 + 0.36443089V + 1.54330901$
Large Bus	$: L_{lb} = -0.00073162V^2 + 0.10371089V + 1.06854641$
Small Truck	$: L_{st} = -0.00205073V^2 + 0.25711696V + 2.90910340$
Medium Truck	$: L_{mt} = -0.00136819V^2 + 0.16318950V + 1.06722744$
Large Truck	$: L_{lt} = -0.00042379V^2 + 0.05886221V + 0.88966832$

where V = Speed.

On the other hand, VSP (vehicle-specific power, kW/ton) has been studied as a research model for estimating fuel consumption. The VSP was proposed by Jimenez-Palacios [17]. The power output of the vehicle has a strong correlation with the fuel consumption. Eq. (1) is the expression of VSP.

$$VSP = \frac{\frac{d}{dt}(KE+PE)+Rolling^{v}+\frac{1}{2}\rho_{\alpha}C_{D}A(v+v_{W})^{2}v}{m} \tag{1}$$

where:

KE = kinetic energy of the vehicle,
PE = potential energy of the vehicle,
$Rolling^{v}$ = rolling resistance of the vehicle suffered,
ρ_{α} = air density,
C_D = drag coefficient,
A = cross-sectional area of the vehicle,
v = speed of the vehicle,
v^{w} = speed of the vehicle against the wind,
m = mass of the vehicle.

However, it is difficult to obtain all these parameters in a real driving environment. Zhou et al. [18] established the modified Eq. (2), which is derived from Eq. (1):

$$VSP = v \times (1.1 \times a + 0.132) + 0.000302 \times v^3 \tag{2}$$

where:

v = speed of the vehicle,
a = acceleration of the vehicle.

The equation only requires speed and acceleration, which are relatively easy to determine. These studies were successful in theoretical settings, but certain limitations need to be taken into consideration, such as poor driving conditions and dangerous driving behaviors.

2.3.2 Study of the relationship between driving behavior and fuel consumption

Son et al. [14] analyzed the effects of different driving style characteristics on fuel consumption. He developed an estimated model between fuel usage and driving behavior variables such as the average vehicle speed, the average RPM, the depth of accelerator pedal depression, the number of uses, and the braking times. In addition, he applied regression analysis on these variables. As a result, he derived a model that can contribute to the calculation of fuel consumption using only driving behavior variables in R^2=0.852. In this study, the depth of the accelerator pedal and the rotational speed of the average RPM had a 78.8% contribution.

Ericsson [15] investigated the characteristics of the major impacts on fuel usage, and defined the driving patterns with an independent measurement method. Sixty-two parameters were analyzed, and extracted into 16 independent driving pattern factors. While calculating the regression analysis of relationship among the fuel consumption and the driving pattern factors, he observed that nine pattern factors have a noticeable effect. His research analyzed that there is an explanatory power of 76% (R^2=0.76) in the relationship between the fuel consumption and the 16 driving pattern factors.

Chi et al. [19] studied effect of driving pattern parameters on fuel economy. He set 13 driving pattern parameters and tried prediction and forecasting using regression analysis. He compared the effects of fuel consumption on the driving pattern parameters of traditional diesel buses and hybrid electric city buses (HEV). In Table 1, the effect shows four grades of the absolute values of standardized B values. The most significant impacts on fuel consumption are the percentage of idle time, average speed, and the percentage of time in speed interval 0–20 km/h. The idle time is a waste of fuel due to the running engine without driving. We can also see that fuel consumption is significantly affected by driving with speed interval 0–20km/h.

Kang et al. [16] developed a model for analysis of fuel consumption according to the type of driving pattern, using the regression analysis in his research into eco-driving activation plans through the analysis of traffic flow and the driving patterns. The results of the analysis of the data of 60 factors and 7 variables including the sudden start, sudden acceleration, sudden stop, sudden deceleration, and sudden turn left/right change are as shown in Table 2. In this study, it was confirmed that a sudden driving pattern had a significant effect on the fuel consumption. It has been analyzed for the large effects that the sudden acceleration and the sudden starting have on fuel consumption, and how the sudden stops and the rapid deceleration have less effect on fuel consumption. However, if the sudden deceleration *p*-value is 0.33, this is not statistically significant at the 0.1 significance level of this study, and it was found that the reliability of the impact on fuel consumption of sudden deceleration decreases.

Table 1. Driving pattern parameters' effects on fuel consumption for diesel bus [19]

Driving pattern parameters	Effects
% of idling time	++++
Average speed	++++
% of time in speed interval 0_20 km/h	++++
Average acceleration	+++
Acceleration root mean square	+++
Average driving speed	++
% of accelerating time	++
Average lengths of microtrips	++
Average deceleration	+
Average time of microtrips	+
Maximum speed	+
% of decelerating time	+
% of time in speed interval 20_40 km/h	+

Table 2. Fuel consumption analysis according to driving type [16]

Model	Unstandardized Coefficients		Standardized coefficient β	*t*	Sig.
	B	Std. Error			
Constant	-17.354	2.690	-	-6.451	0.000
Sudden Start	10.248	0.527	0.232	19.454	0.000
Sudden Acceleration	14.256	0.647	0.412	22.022	0.000
Sudden Stop	1.390	0.463	0.046	3.001	0.004
Sudden Deceleration	0.524	0.540	0.017	0.970	0.337
Sudden lane change	4.007	0.368	0.225	10.876	0.000
Continuous sudden lane change	6.729	0.516	0.275	13.030	0.000
Sudden turn left/right	3.376	0.928	0.052	3.639	0.001

In these studies, it is possible to estimate the fuel consumption through the driving pattern of the driver. Further in the present study, we would like to find that the formula could be used to calculate the driving pattern factor of the driver and to estimate the fuel consumption by using the DTG data. We can get a great deal of data easily because, In Korea it is obligatory for the commercial vehicles to have a device mounted on it. It is possible that the fuel consumption economy can be estimated by DTG data and can be used to help the drivers or the entire transportation companies. We have tried to review the DTG data in Section 3.

3. Analytical Environment of DTG

3.1 Introduction of DTG Device

As the movement to reduce the traffic accidents in Korea, DTG devices have been legally required to be mounted on the commercial vehicles since June 2014. The DTG records a number of data, such as the vehicle's GPS coordinates, speed, acceleration, RPM, etc., at one-second intervals over 6 months period, and it must be submitted periodically to the Korea Transportation Safety Authority. DTG is a device that records the operational state of the vehicle in combination with clock, speedometer, and odometer. In the past, analog-type tachographs have been stored, and a mechanical disc-shaped recording paper was used. However, the device changes the data to a digital type to record the status of the vehicle’s operations accurately.

3.2 Introduction of DTG Data

DTG data is composed of the Header and Body parts. In the Header part, general vehicle information is stored. In the Body part, detailed data of DTG is stored in the record unit. Table 3 shows a sample of the data recorded on the DTG Body part. The size of one-day driving data per car is about 3.4 Mbytes in text format. The size over 3 months is about 300 Mbytes. Each transport company needs to transfer the DTG data of each vehicle in a 3-month period to the Korea Transportation Safety Authority [20].

Table 3. DTG data body sample

Date/Time	Speed	RPM	Brk	GPS_X	GPS_Y	Dir	AccX	AccY
140101/05575700	61	1240	0	127002710	37607548	288	0.9	1.2
140101/05575800	61	1390	0	127002150	37607685	290	0.1	0
140101/05575900	61	1390	0	127002150	37607685	292	-1.6	0.1
140101/05580000	62	1400	0	127002150	37607685	292	0.1	0.1
140101/05580100	62	1420	0	127001606	37607863	293	-0.5	0
140101/05580200	63	1420	0	127001606	37607863	293	-0.3	-0.1
140101/05580300	63	1430	0	127001606	37607863	293	1.7	0
140101/05580400	63	1450	0	127001055	37608050	293	-1.6	-0.1

3.3 Introduction of DTG Dangerous Driving Behavior Statistical Service

In most of the companies, DTG data is used as statistical data for dangerous driving such as over speed, sudden driving, idling period, and so on. TS (the Korea Transportation Safety Authority) collects the DTG data of commercial vehicles from all the transportation companies in Korea. TS offers a dangerous driving statistics service to drivers. Companies and the drivers check the statistics service at its homepage [21]. Table 4 can be used for dangerous driving statistics on TS, and it defines 11 types of dangerous driving behavior and provides statistics.

Table 4. Definition of dangerous driving acts of TS [21]

Dangerous driving behavior	Definition
Overspeed type	
Overspeed	Driving more than 20 km/h above the speed limit of the road
Long-term overspeed	Driving more than 20 km/h from the speed limit of the road and keep the speed more than 3 minutes
Sudden acceleration type	
Sudden acceleration	Accelerating driving more than 11 km/h per second
Sudden start	Starting from the stopped state and accelerating, increasing velocity by more than 11 km/h per second
Sudden deceleration type	
Sudden deceleration	Decelerating driving more than 7.5 km/h per second
Sudden stop	Decelerating driving more than 7.5 km/h per second, and speed becomes "0"
Sudden rotation	
Sudden right rotation	Speed is 15 km/h or more and rapidly rotated to the left (range 60°–120°) in 2 seconds
Sudden left rotation	Speed is 15 km/h or more and rapidly rotated to the right (range 60°–120°) in 2 seconds
Sudden U-turn	Speed is 15 km/h or more, rapidly U-turning to the left or right (range 160°–180°) in 2 seconds
Sudden course change type	
Sudden overtaking	Overtaking with changing lane the car to the right or left (range 30°–60°) while accelerating 11 km/h per second or more
Sudden course change	Accelerating or decelerating by changing the car to the right or left (range 15°–30°) over 30 km/h speed

The Ministry of Land, Infrastructure and Transport in Korea has analyzed its application in actual cases, and the effect of having a DTG mounted. It stated that “there are companies that have reduces 50% of traffic accidents or reduces less than half insurance rates compared to the pre-mount” [22]. However, the statistics service of fuel consumption can only be provided if the driver directly enters their fuel consumption, because research for estimating the fuel consumption with the DTG data alone has not yet progressed.

4. Fuel Consumption Estimation Processing

This section provides our approach of fuel consumption estimation processing. The fuel consumption estimation processing flow chart is shown in Fig. 1. The collected DTG data and the OBD-II fuel consumption data are used to analyze the fuel estimation. Those data are processed by the Hadoop MapReduce (MR) mechanism to maximize the parallel processing. The results from the MR processing are the immediate result sets of analysis and driving patterns. In the next step with those diving pattern data, the regression function in R is used to generate the fuel estimation results.

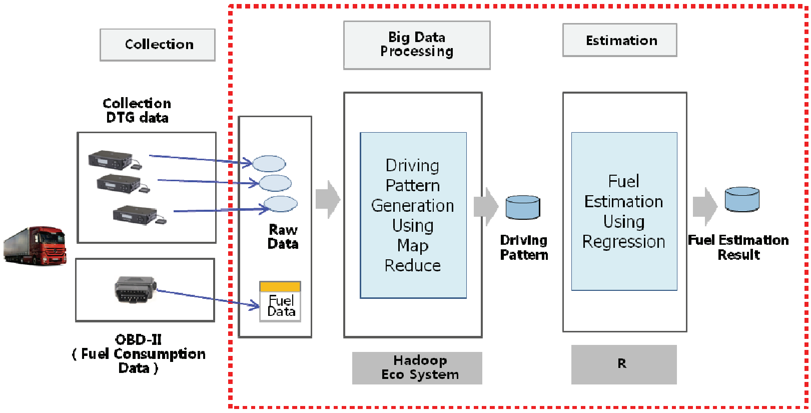

Fig. 1. Fuel consumption estimation processing flow chart.

In order to estimate the more accurate fuel consumption data, we have analyzed the actual fuel consumption data from the mounted OBD-II devices. We have generated a model that has the real fuel consumption data as a target variable and the driving pattern data as independent variables. Based on the analysis, it is possible to provide economic driving indices to transportation companies by estimating the fuel consumption only by using the DTG data of vehicles.

4.1 DTG Big Data and Mileage Data

4.1.1 DTG big data collection

In this paper, the data used for the analysis is the actual data provided by the commercial vehicle traffic monitoring service in Korea [23]. The vehicle that carries the DTG device is a kind of a cargo truck or bus. We collects this data, which is sent to the server through a mobile communication using a 3G modem and stored in another space. We collect second-by-second DTG data and transmit to the server via modem in every 5 minutes. Data from 4,605 vehicles of the total 223 companies, over 5 months period and 416,238 journeys were used. Journey is defined as data about a single vehicle that has traveled for 1 day, and one file is generated in the unit Journey. An average of 3.4 Mbytes per file is generated, and each company has an average of 21 vehicles. The total data size is about 1.3 Tbytes. Commercial vehicles with mounted DTG such as trucks have the characteristic of the data that has been driving on expressway, national roads, and prefectural roads rather than roads in cities.

Table 5. One-second unit fuel consumption data

Date/Time	Total cumulative fuel consumption	Daily cumulative fuel consumption (mL)
20140201/122556	148409067	0
20140201/122557	148409135	68
20140201/122558	148409150	83
20140201/122559	148409166	99
20140201/122600	148409181	114
20140201/122601	148409196	129
20140201/122602	148409212	145

4.1.2 Fuel consumption data

Eight vehicle's fuel consumption was measured using OBD-II devices. Over 5-month period, 1,559 fuel data points were collected (Feb–Jun 2014). To observe the daily fuel consumption, we let a data file be the 1-day fuel consumption data of one car. Fuel efficiency is determined by measuring the total distance traveled and the total operating time of the day, the fuel consumption was calculated as the mileage per liter. We used a total 493 data, excluding garbage. At least 29 files and at most 67 were used from each vehicle with encoded car numbers. Using 1-second unit serialized fuel consumption data (Table 5), we can recognize which variables affected the fuel consumption.

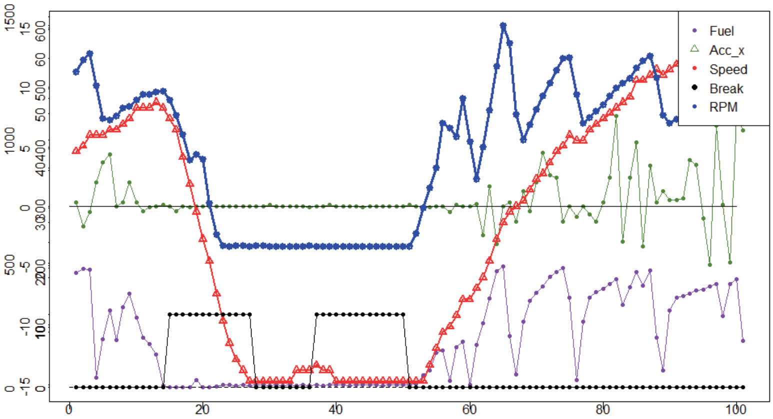

Fig. 2. DTG data graph that contains the fuel consumption.

Fig. 2 is a sample graph of 100 seconds of speed, RPM, break signal, acceleration X, and fuel consumption using R visualization package [24]. The red line shows the speed, the blue shows the RPM, the black shows the brake signal, the green shows the acceleration X (long axis), the purple shows the fuel consumption, and this displays the motion during the 100 seconds. As for this vehicle DTG data, correlations between fuel consumption and speed are 0.483, and RPM is 0.532, and break signal is 0.243. RPM was found to have a slightly higher correlation than the other variables.

4.2 Pattern Extraction to the Fuel Consumption Measurement using DTG Big Data

4.2.1 Driving pattern definition

Through the prior research detailed Section 2.3, we have defined the driving patterns by extraction using the DTG big data as in the following Table 6. We extracted this driving pattern by statistically analyzing the DTG data. Variables 1–5 are the daily general driving statistics, variables 6–11 are the dangerous driving behavior statistics such as overspeed and sudden driving behaviors. Variable 12 shows the idling when the speed is zero and the RPM is greater than zero. Variable 13 is the average RPM, variable 14 is a number of times the brakes have been hit. Variable 15 is the average break time when the driver stepped on the brake. Variables 16–23 are the average and standard deviation of the speed and the RPM. Variable 16 is the average of acceleration and variable 17 is the standard deviation of acceleration. Variables 18–19 are the average and standard deviation of deceleration. Variables 20–21 are the average and standard deviation of the RPM change, variable 22 is the standard deviation of speed, and variable 23 is the standard deviation of the RPM. Variables 24–30 are each rate of the seven ranges of speed such as 0–15 km/h. That is, the speed_0_15 value will be 100% when the driver is traveling at a speed within the range 0–15 km. Variables 31–38 are the percentages of each section the number of RPM. In total, 38 variables were used as independent variables.

Table 6. Driving pattern using DTG data

No	Variable	Description
1	DayDist	Total driving distance of each trip
2	avgSpd	Average speed of each trip
3	highestSpd	Highest speed of each trip
4	highRPM	Highest RPM of each trip
⋮	⋮	⋮
13	avgRpmCnt	Average RPM of each trip
14	totBreakCnt	Total break hit count of each trip
15	avgBreakTime	Average break time during each brake hit
16	avg_Accel	Average acceleration
17	SD_Accel	Standard deviation of acceleration
18	avg_Decel	Average deceleration
19	SD_Decel	Standard deviation of deceleration
20	avg_DRpm	Average RPM of increase or decrease
21	SD_DRpm	Standard deviation RPM change of RPM increase or decrease
22	SD_Speed	Standard deviation of speed
23	SD_Rpm	Standard deviation RPM
24	speed_0_15	Speed 0–15 km rate of count
⋮	⋮	⋮
30	speed_110_200	Speed 110–200 km rate of count
31	rpm_0_5	RPM 0–500 rate of count
⋮	⋮	⋮
38	rpm_35_99	RPM 3500 rate of count

4.2.2 Pattern extraction using Hadoop MapReduce mechanism

For processing of huge scaled DTG sensing data, we apply the big data processing method with parallel processing. In this study, we designed and implemented DTG big data analysis using the Hadoop MapReduce technique. The MapReduce mechanism establishes the key-value setting. The Mapper process operates grouping of each journey unit. The value is generated from the journey sequence of DTG data such as speed, rpm, brake signal, direction, acceleration, and GPS coordinates. The Algorithm 1 shows the Mapper process, which generates the keys and the values.

Algorithm 1. Mapper of driving pattern

```
Input : ( key: offset in bytes; value: text of a record )
Output : ( key' : key of the record, value': value of the record )

1. key' = CarNum + Date
2. value' = speed + rpm + brakesign + dir + acc + gps_x + gps_y
3. output ( key', value' )
```

The Reducer process applied the statistical operation. The generated values are the driving patterns which are defined in the previous Section l as DayDist, average Speed, highest Speed, SD_Speed, speed_0_15, …, etc.

The Algorithm 2 shows the algorithms of the Reducer process.

Algorithm 2. Reducer of driving pattern

```
Input : ( key: key of the group records, values: values of the group records )
Output : ( key' : key of the record, value': value of the driving pattern record )

1. calculate each driving pattern by the group values of input key ( average, standard deviation, each
   rate of ranges, etc )
2. value' : DayDist + avgSpd + highestSpd ....  rpm_35_99
3. output ( key, value' )
```

To produce a statistic and driving pattern, we have suggested and developed a method using big data processing. Table 7 is an example of pattern data.

Table 7. An example of driving pattern data

CarNum	Date	Dist	Avg Speed	High Speed	…	Avg Accel	SD Accel	Avg Deccel	SD Deccel	…	s_0_15	s_15_30
GG81F00293	140501	301	44	117	…	1.500	0.981	- 1.656	1.143	…	36.15	7.02
GG84H00113	140501	169	36	97	…	1.359	0.782	- 1.543	1.033	…	39.78	7.52
GG85I08016	140501	231	34	114	…	1.682	1.085	- 1.729	1.149	…	40.68	10.46
GG85I08031	140501	462	60	114	…	1.612	1.086	-1.709	1.386	…	17.77	6.63
GG90H02566	140501	392	46	105	…	1.371	0.671	-1.637	1.069	…	19.84	9.96
GG93H01474	140501	22	7	65	…	1.666	0.921	-2.004	1.428	…	74.14	18.87
GG93H01475	140501	158	17	105	…	1.527	0.868	-1.821	1.256	…	63.66	14.86
GG93H01477	140501	45	13	82	…	1.577	0.790	-2.220	1.455	…	59.98	21.68
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

4.3 Essential Data Analysis of the Driving Pattern

4.3.1 Comparison between the percentage of dangerous driving behavior statistics and the mileage

Fig. 2 is the scatter plots of dangerous driving pattern and the mileage using the plot() R function [25]. The dangerous driving behavior variables include the excessive speeding, the dangerous speeding, the sudden acceleration, the sudden deceleration, the sudden starting, the sudden stopping, and the idling count. In Fig. 3, it can be seen that the correlation between the fuel consumption and the dangerous driving behavior was not high. It was estimated that the number of dangerous driving patterns affecting the regression coefficient would not be significant, when considering the reasons. The criteria for calculating the statistics using the DTG driving pattern data for the regression was not sufficient to significantly impact the fuel consumption. In order to improve this, it is necessary to reduce the criteria for statistical calculations.

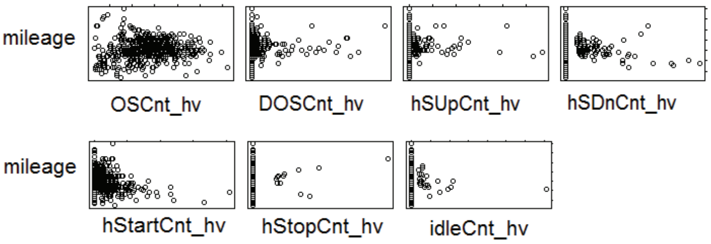

Fig. 3. Scatter plots of the relationship between dangerous driving pattern and mileage.

4.3.2 Comparison between the percentage of speed ranges and the mileage

In this subsection, we are going to compare the impact between the speed range and the mileage. Fig. 4 consists of scatter plots of the relationships between each speed range and mileage. The x-axis is each speed range (percentage) and y-axis is mileage level. It can be seen that as the ratio of the speed_0_15 section is high, the fuel consumption is reduced. In the rest, it can be seen that the fuel economy is also increased gradually as the mileage increases. We can derive that if the driver uses more low-speed driving, their fuel consumption will increase.

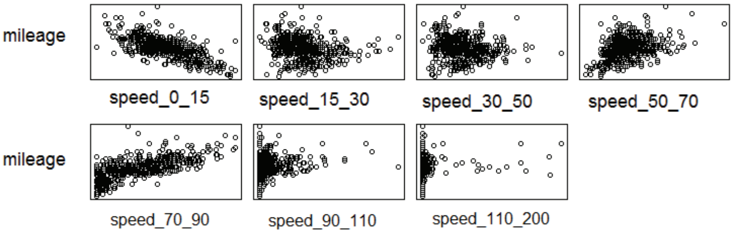

Fig. 4. Scatter plots of the relationship between speed range and mileage.

4.3.3 Comparison between the percentage of RPM range and the mileage

Fig. 5 is the scatter plots of RPM range and mileage. The x-axis is each RPM range (percentage) and y-axis is mileage level. In the RPM range 500–1500, mileage becomes lower as the RPM count increases, and in the RPM range 1000–1500 it becomes higher as the RPM count increases. In other ranges, the mileage relationship is less evident.

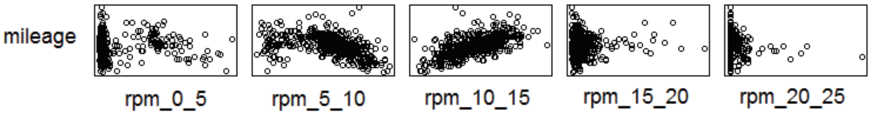

Fig. 5. Scatter plots of the relationship between RPM range and mileage.

4.3.4 Comparison between the mileage and the statistical data

Fig. 6 is scatter plots of the relationship between the statistics data and mileage. The x-axis is each statistics data count (number of cases) and y-axis is mileage level. It is possible to extrapolate a proportional relationship between the mileage, the SD_ Speed and the avg_Rpm.

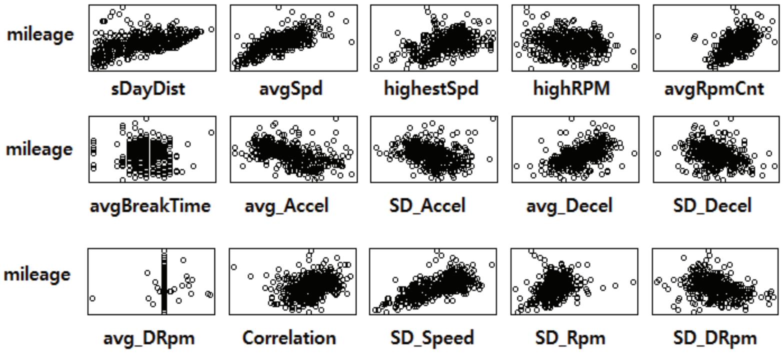

Fig. 6. Scatter plots of the relationship between statistics data and mileage.

5. Analysis Result

We work on big data analysis for the fuel consumption estimation in this section. We use a regression analysis using the driving pattern which is shown and generated in Section 4.

5.1 Fuel Economy Formula Analysis using Driving Pattern

We have selected the variables using the regsubsets() R function [26] for the model selection by exhaustive search of the regression analysis as shown in Fig. 7. The highest adjR^2 variables selected were highestSpeed, avgRpm, SD_Accel, avg_Decel, SD_Decel, SD_Speed, SD_Rpm, and speed_0_15.

Table 8 shows the result of regression analysis in R. The adjusted R^2 is 0.732. It has an explanatory power of 73.2%. From the results, this can be interpreted as representing the main variables used. The B value of SD_Accel (the standard deviation of acceleration) is 0.7179. These could suggest that the fuel efficiency will improve as the standard deviation of acceleration is high. SD_Decel (the standard deviation of deceleration) and the SD_Speed (the standard deviation of speed) could also suggest that the fuel efficiency will improve, as these values are high.

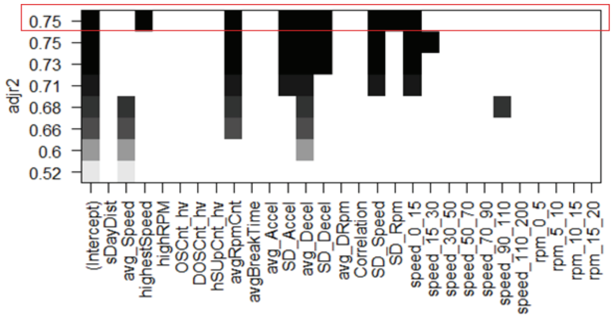

Fig. 7. The results of regsubsets() R function.

Table 8. The result of regression analysis in R

Variable	Coefficient	Standard error	*t*	Pr(>|*t*|)
(Intercept)	6.502789	0.356276	18.252	<2e-16 ***
highestSpeed	-0.007511	0.001519	-4.946	1.05E-06 ***
avgRpmCnt	-0.003427	0.000287	-11.941	<2e-16 ***
SD_Accel	0.717953	0.108386	6.624	9.32E-11 ***
avg_Decel	-0.175657	0.018858	-9.315	<2e-16 ***
SD_Decel	0.580536	0.141853	4.093	5.00E-05 ***
SD_Speed	0.056198	0.004837	11.618	<2e-16 ***
SD_Rpm	-0.002169	0.000437	-4.959	9.82E-07 ***
speed_0_15	-0.033139	0.002066	-16.042	<2e-16 ***

Significant codes: *** $p<0.001$.

As in the SD_Accel and the SD_Speed, unlike the other studies, the fuel economy increases as the standard deviation of the speed increases in this set of DTG data. It can be assumed that the result of the properties of the DTG data of commercial vehicles such as trucks that are driven frequently on highways nationwide, and national roads are reflected. In other words, there is a tendency for a relatively high mileage and high standard deviation of speed when vehicles often drive on highways. Fig. 8 show a scatter plot of the relationship between the average speed and the standard deviation of the speed, and one of the relationship between the highest speed points per vehicle and the standard deviation of the speed. It means that the standard deviation of the rate is high when the speed of the vehicle is generally high.

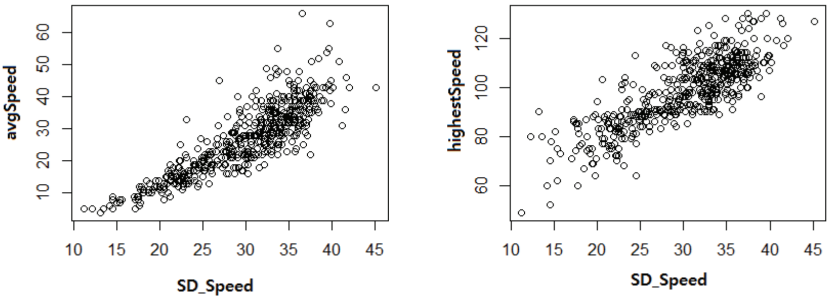

Fig. 8. A scatter plot of the average speed and standard deviation of speed, and one of the highest speed and standard deviation of speed.

In order to build the fuel consumption efficiency, the coefficient values derived through regression analysis as shown in Table 8 are applied to the fuel consumption estimation formula as follows.

$$
\begin{aligned}
\textit{Predicted Mileage} = 6.5028 &- 0.0075 \times \textit{Speed}_{(highest)} \\
&-0.0034 \times \textit{RPM}_{(average)} \\
&+0.7180 \times \textit{Acceleration}_{(Standard\ Deviation)} \\
&-0.1757 \times \textit{Deceleration}_{(average)} \\
&+0.5805 \times \textit{Deceleration}_{(Standard\ Deviation)} \\
&+0.0562 \times \textit{Speed}_{(Standard\ Deviation)} \\
&-0.0022 \times \textit{RPM}_{(Standard\ Deviation)} \\
&-0.0034 \times \textit{Speed}_{(rate\ of\ range\ 0\sim15km)}
\end{aligned}
$$

5.2 Model Validation

The regression model was evaluated in three different approaches: comparing the predicted values to the observed values, considering the error rates, and the residual analysis.

5.2.1 Prediction verification

As the first verification step of analysis, we use the fuel consumption estimation formula derived in the previous subsection of fuel economy formula analysis.

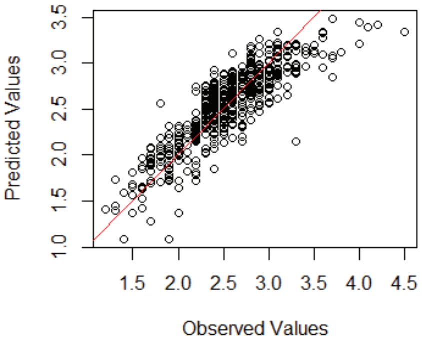

Fig. 9. The plot chart of the relationship between the observed mileage and the predicted values.

Fig. 9 is a plot chart of the relation between the observed mileage and the predicted values by using the formula derived. The correlation coefficient of the groups of two values is 0.8581. We calculated and drew in the density histogram graph of the enlarged view of all 4,605 vehicles by the fuel consumption estimation formula as shown in Fig. 9. By filtering with the same range of variables used in the regression equation for fuel consumption estimation, it was estimated with the data of 191,221 of the total 416,238 cases. Most of the values were located between 0–6 mileages. However, some of the data was located on a negative value (1006 values, 0.5%).

In Fig. 10, the density graph scatters over even negative values of mileage from the estimation formula. It is because we selected eight vehicles to measure the fuel consumption in analysis phase, and we applies to more generalized cases of 4,605 vehicle. The fuel mileages vary greatly depending on the type and weight of vehicles. Therefore, we know that in order to expand the model to all commercial vehicles, it is necessary to consider additional variables of the vehicles such as the load, the vehicle type, and so on, to improve the accuracy of the formula.

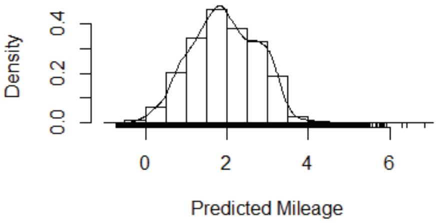

Fig. 10. The density graph of the estimated fuel economy of the 4,605 vehicles.

5.2.2 Error rate

To find the accuracy of the fuel consumption estimation formula and determine if it is acceptable, the error rate of the fuel consumption estimation model was calculated with the following formula:

$$S = \frac{|\,Fuel_{real} - Fuel_{model}\,|}{Fuel_{real}} \times 100\% \tag{3}$$

where:

Fuel$_{real}$ = real fuel consumption,

Fuel$_{model}$ = fuel consumption calculated by model.

$$S = \left(\sum_{i=1}^{493} S_i\right) / \, 493 \tag{4}$$

where:

S$_i$ = error rate of the fuel consumption second-by-second data i.

The overall error rate of all samples was 7.68%, and the maximum error rate of a single trip was 42.6%. Fig. 11 shows the density histogram plot of the error rate.

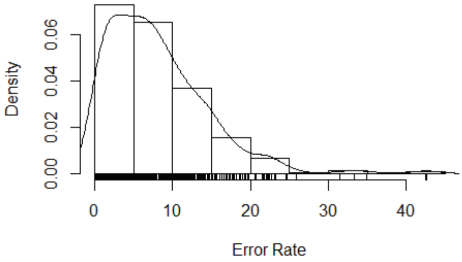

Fig. 11. The density histogram plot of error rates.

The proportion of samples with an error rate of less than 10% was 69.4%. The proportion of samples with an error rate of less than 20% was 95.6%. It shows that these error rates are acceptable to have the estimation model confidential.

5.2.3 Residual analysis

To observe fitting evaluation of the model of the regression equation derived, we worked residual scaling shown in Fig. 12.

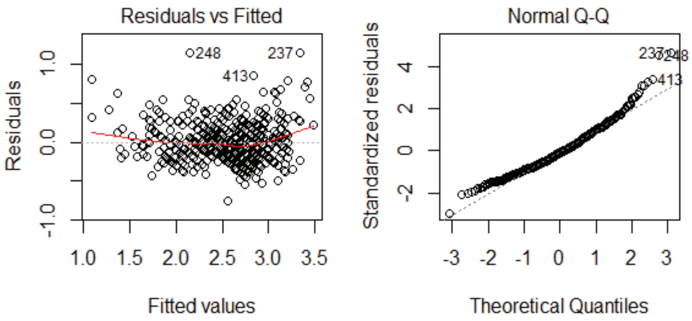

Fig. 12. Model evaluation charts.

The charts indicate that residuals locate the difference of the actual data values and fitted values adapted to the regression on the x-axis and y-axis. It shows that the linear slope is close to zero and the variance of residuals follows the normal distribution constant. We admit the model acceptable as a relatively stable model. As another chart for residual analysis, the Normal Q-Q has the purpose of verifying that the residuals are normally distributed. When it has a linear relationship, we can see that it is close to normal distribution, resulting in residual verification for the model validation.

The results from Section 5 also have shown a significant result of DTG big data processing and analysis. Beyond using the simplified data of speed and vehicle data, the enhanced information including the standard deviation, ranges of speed, and driving patterns has provided the meaningful results with accumulated vehicle driving data, yielding the appropriate big data analysis for fuel consumption.

6. Conclusion

6.1 Summary

In this paper, we have researched to estimate the fuel consumption mileage using the driving patterns extracted from the DTG big data of commercial vehicles. Compared to the related researches that require fuel consumption data, we analyze the data correlation of DTG data, which contains driving patterns, especially actual driving logging data of commercial vehicles. Also we apply the parallel processing of Hadoop MapReduce mechanism for the benefit of fast processing on DTG big data.

We have derived an analytical fuel consumption formula to estimate the fuel mileage using DTG data variables through a linear regression. In Section 5.1, we derived the formula of the fuel mileage estimation by regression analysis using 38 driving pattern variables. Among those variables, we figured out that 8 major variables contribute to the fuel consumption estimation: Acceleration (standard deviation), Deceleration (standard deviation), Deceleration (average), Speed (standard deviation), Speed (rate of range 0–15 km), Speed (highest), RPM (average), RPM (standard deviation) in order of influence of variables.

To find the deriving model of linear regression valid, we compared the actual data with the predicated one, confirmed if error rates are acceptable, and observed the fitting scale of residuals. The fuel consumption estimation formula was derived from driving pattern using DTG big data of actual commercial vehicles running on highway roads in Korea. Compared to other research work, we have achieved the big data analysis for the fuel consumption estimation with the collection of actual driving data and the OBD-II fuel consumption data. Another major research contribution is to analyze a large amount of actual driving data of commercial vehicles from the real industrial field, which has a ripple effect on the driving fuel estimation and validation. This study is able to be used as eco-driving service for most commercial vehicles mounted DTG devices in Korea.

6.2 Future Work

Though we derived a fuel mileage estimation formula with only DTG data, because it was not calculated for all vehicles, it is necessary to derive a more general formula by distinguishing the type of vehicle load or each vehicle type. Then, we thought that it could calculate a more accurate fuel mileage formula by deriving other variables such as road condition, driver type, and other environmental data that can be collected. In addition, an aim of our future research is to study eco-routing technology that includes which road costs the least amount of fuel, using the estimated fuel consumption cost of each road.

Acknowledgement

This research was supported by the Ministry of Science, ICT and Future Planning, Korea, under the IT/SW Creative research program supervised by the National IT Industry Promotion Agency (NIPA-2013-H0502-13-1071). Also it is a result of a study on the “Leaders in Industry-University Cooperation (LINC)” project, supported by the Ministry of Education, Korea.

References

[1] K. Boriboonsomsin, A. Vu, and M. Barth, “Eco-driving: pilot evaluation of driving behavior changes among US drivers,” University of California Transportation Center, Riverside, CA, 2010.

[2] G. A. Klunder, K. Malone, J. Mak, I. R. Wilmink, A. Schirokoff, N. Sihvola, et al., “Impact of information and communication technologies on energy efficiency in road transport: final report,” TNO, Delft, The Netherlands, 2009.

[3] Korea Ministry of Land, Infrastructure and Transport, "Mandatory to mount DTG on all commercial vehicles," 2010 [Online]. Available: http://www.molit.go.kr/USR/NEWS/m_71/dtl.jsp?id=155552574.

[4] J. Kang, Y. Kim, U. Lim, and M. Jun, “An improved vehicle data format of digital tachograph,” *Journal of the Korea Society of Computer and Information*, pp. 77-85, 2013.

[5] M. Barth and K. Boriboonsomsin, “Energy and emissions impacts of a freeway-based dynamic eco-driving system,” *Transportation Research Part D: Transport and Environment*, vol. 14, no. 6, pp. 400-410, 2009.

[6] J. Park, *Review of Eco-Driving Policy in Advanced Countries and Its Implication*. Seoul: The Korea Transport Institute, 2009.

[7] N. Jeon, H. Ham, K. Jeong, and H. Lee, "Development of eco-driving monitoring algorithm based energy efficiency," in *Proceedings of the Korea Society of Automotive Engineers (KSAE) Spring Conference*, 2012, pp. 942-948.

[8] H. Rakha, K. Ahn, and A. Trani, "Development of VT-Micro model for estimating hot stabilized light duty vehicle and truck emissions," *Transportation Research Part D: Transport and Environment*, vol. 9, no. 1, pp. 49-74, 2004.

[9] H. Rakha, H. Yue, and F. Dion, "VT-Meso model framework for estimating hot-stabilized light-duty vehicle fuel consumption and emission rates," *Canadian Journal of Civil Engineering*, vol. 38, no. 11, pp. 1274-1286, 2011.

[10] M. Won, G. Gang, and J. Kim, "A estimation model of the fuel consumption based on the vehicle speed pattern," *Journal of Korean Society of Transportation*, vol. 29, no. 4, pp. 65-71, 2011.

[11] *The 5th Amendment of Transportation Facility Investment Evaluation Guidelines*. Seoul: Ministry of Land, Infrastructure and Transport, 2013.

[12] G. Scora and M. Barth, "Comprehensive modal emissions model (CMEM) version 3.01 user guide," Centre for Environmental Research and Technology, University of California, Riverside, CA, 2006.

[13] S. Vallamsundar and J. Lin, "Overview of US EPA new generation emission model: MOVES," *ACEEE International Journal on Transportation and Urban Development*, vol. 1, no. 1, pp. 39-43, 2011.

[14] J. Son, M. Park, H. Oh, J. Lee, and T. Lee, "Age and Gender difference in fuel efficiency on highway driving," in *Proceedings of the Korea Society of Automotive Engineers (KSAE) Spring Conference*, 2013, pp. 264-268

[15] E. Ericsson, "Independent driving pattern factors and their influence on fuel-use and exhaust emission factors," *Transportation Research Part D: Transport and Environment*, vol. 6, no. 5, pp. 325-345, 2001.

[16] K. Kang, J. Oh, J. Park, and N. Sung, *Eco-Driving based on an Analysis of Driving Patterns and Traffic Flow*. Seoul: The Korea Transport Institute, 2010.

[17] J. L. Jimenez-Palacios, "Understanding and quantifying motor vehicle emissions with vehicle specific power and TILDAS remote sensing," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, 1998.

[18] X. Zhou, J. Huang, W. Lv, and D. Li, "Fuel consumption estimates based on driving pattern recognition," in *Proceedings of International Conference on Green Computing and Communications (GreenCom), 2013 IEEE and Internet of Things (iThings/CPSCom), and IEEE International Conference on and IEEE Cyber, Physical and Social Computing*, Beijing, China, 2013, pp. 496-503.

[19] M. Chi, H. Wang, M. Ouyang, "Effect of driving pattern parameters on fuel-economy for diesel and hybrid electric city buses," in *Proceedings of International Electric Vehicle Symposium and Exhibition,* Goyang, Korea, 2015.

[20] Korea Transportation Safety Authority, "Digital Tachograph Analysis System," [Online]. Available: http://etas.ts2020.kr.

[21] Korea Transportation Safety Authority, "Dangerous driving behavior criteria," 2015 [Online]. Available: http://etas.ts2020.kr/etas/frtl0401/pop/goList.do.

[22] Korea Ministry of Land, Infrastructure and Transport, "The effect of having a DTG mounted," 2011 [Online]. Available: http://www.molit.go.kr/USR/policyTarget/m_24066/dtl.jsp?idx=311.

[23] DTG monitoring service [Online]. Available: http://tacho.gtrac.co.kr.

[24] W. Cho, Y. Lim, H. Lee, M. K. Varma, M. Lee, and E. Choi, "Big data analysis with interactive visualization using R packages," *Proceedings of the 2014 International Conference on Big Data Science and Computing*, Beijing, China, 2014.

[25] The R Foundation for Statistical Computing [Online]. Available: https://www.r-project.org/.

[26] T. Lumley, Package 'leaps' [Online]. Available: https://cran.r-project.org/web/packages/leaps/index.html.

Wonhee Cho http://orcid.org/0000-0001-9087-5545

He received B.S. and M.E. degree in Computer Science & Engineering from Inha University in 1990 and 1992, respectively. He obtained his Ph.D. in the Graduate School of Business IT at Kookmin University in 2016. He had worked as a research engineer of LBS/Telematics area for 21 years at SK in Korea. He is an adjunct professor at Korea Soongsil Cyber University and a visiting scholar at University of Southern California. His current research interests include IoT and vehicle trajectory big data analysis.

Eunmi Choi http://orcid.org/0000-0002-9743-2437

She is a Professor in Kookmin University, Korea. Her current research interests include big data infra system and analysis, cloud computing, intelligent system, information security, parallel and distributed system, and SW architecture and modeling. Professor Choi received and M.S. and Ph.D. in Computer Science from Michigan State University, USA in 1991 and 1997, respectively, and B.S. in Computer Science from Korea University in 1988 with the top student award. Since March 1998, she had worked as an assistant professor in Handong University, Korea, before joining Kookmin University in 2004. She is the Head of Distributed Information System & Cloud Computing Lab., in Kookmin University.