IJESRT
INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

APPLICATION OF DATA MINING METHODS FOR ANALYZING OF THE FUEL CONSUMPTION AND EMISSION LEVELS

Simge Deniz [*], Hadi Gokcen [*], Gholamreza Nakhaeizadeh
[*] Department of Industrial Engineering, Gazi University, Turkey
Department of Economics and Management, Karlsruhe Institute of Technology, Germany

DOI: 10.5281/zenodo.160871

ABSTRACT

This paper is aimed to investigate application potential of data mining in automotive industry. Most important usage and driving parameters, which effects fuel consumption and emission level of passenger cars identified and classified by using data mining methods. A dataset created by combining Euro 6 data of passenger cars has been analyzed using different tools of SPSS, such as; descriptive statistics, correlations, regression and etc. Results have been compared and effecting parameters have been derived by segmentation algorithms aiming better results by categorizing variables for upcoming analysis. The importance of each parameter has been evaluated to predict its contribution on fuel consumption by data mining technics. Therefore, it will be possible to build optimal control strategies for fuel efficiency for future cars, such as; electric, connected and automated vehicles.

Adopted data mining technics in this study are classification algorithms, such as; neural networks, Bayesian networks and C5.0 algorithm as well as segmentation algorithms (e.g. K-means and Two-step) targeting foreseeability and simplicity. Application of those technics by Clementine 12.0 has shown that weight and engine capacity of passenger cars were the most important parameters in fuel consumption, respectively. Depending on the evaluation of the performance of those methods by Evaluation Node of Clementine, it has been found that C5.0 was the most efficient method in prediction of fuel consumption among others. However, the evaluation charts (Gain, Profit, ROI, etc.) have shown that neural network could have better results in prediction in some conditions.

KEYWORDS: Data Analytics, Big Data, Fuel Efficiency, Emissions, Segmentation, Performance Evaluation.

INTRODUCTION

There are many key trends which will impact automotive manufacturers, wholesalers, retailers, customers and drivers over the next years. According to KPMG's 2014 Global Automotive Executive Survey, the rising economic power of the emerging markets remains the central growth force for the next 10 years, the industry become more global and automakers are obliged to use flexible, modular platforms [1].

The globalization of markets, resources and knowledge require product development companies to be highly responsive to customer needs and to environmental changes. In the automotive industry, the major challenges of limiting carbon dioxide (CO_2) emissions while delivering high quality products to a growing number of customers in expanding and highly heterogeneous markets require very efficient and powerful tools and methods to capture customer needs and also to gather performance data and statistics to improve product development. Since the volume of data collected from fleets of vehicles using telematics services can be very high, it is important to design the systems and frameworks in a way that is highly scalable and efficient. This can be described as a Big Data challenge in an automotive context. [2].

The selection of segmentation techniques has become more important due to the fact that the developments in information and communication technologies, especially database management systems and data mining have changed the way of marketing. The vast availability of data and the inefficient performance of traditional statistical techniques (or statistics-oriented segmentation tools) on such voluminous data have stimulated researchers to find effective segmentation tools in order to discover useful information about their markets and customers. Thus, knowledge discovery (KD) and data mining (DM) have been seen as a solution to this problem [3].

According to Harvard Business Review (2016), executives have to clearly define business problems and what the questions are that analytics can answer in order to pursue an effective analytics strategy. If executives don't do this, they risk getting back data that sends the organization in the wrong direction. A type of artificial intelligence that uses algorithms that iteratively learn from data, machine learning can surface insights without being explicitly programmed where to look for them. It makes it easy to crunch massive amounts of data. The speed to insight that machine learning offers can help companies act strategically on the data they have, homing in on the insights with impact, allowing executives to make informed decisions.

The automobile is reliant upon petroleum to fuel our transportation needs. This dependence presents a challenging energy and environmental problem, as the transportation sector is responsible for two-thirds of total petroleum consumption and a third of the nation's carbon emissions. Amid growing concerns over energy security, and the impacts of global climate change, Congress is debating legislative proposals to increase the fuel economy of new passenger vehicles over the next two decades [5].

Considering the efficiency and the intelligence of transportation, the fuel efficiency in vehicle systems gains importance and appears as a possible improvement case in which Big Data can be used as a source for automotive industry. Today, well-known automotive companies tend to produce more economical, environment friendly, technological and personalized cars in context of autonomous driving, electric vehicles and etc. Understanding the importance degrees of usage parameters of vehicles on fuel consumption and emissions will help producers to achieve desired goals and provide drivers eco-driving experiences.

In this paper, fuel consumptions and emission levels of passenger cars have been analyzed by creating a dataset as a combination of Euro 6 and technical data structure of passenger cars. Application potential of big data mining in the automotive industry has been investigated to increase fuel efficiency and to reduce emissions and noise level. Therefore, it will be possible to build optimal control strategies for fuel efficiency.

In this study, data analytics and machine learning methodologies which are happen to use on huge amounts of data (e.g. Big Data) has been investigated. By using the results of this study, different data analytics tools can be compared based on different scenario types. The results of this work will help improving of vehicle design phase in term of vehicle's environmental certification. On the other hand, embedding the results of this paper into mathematical models will improve vehicle routing problems and simulation-based evaluations. In this study, application potential of big data mining has been studied. Data analytics and machine learning methodologies which are happen to use on big amounts of data has been investigated in order to extract variable importance on real-world fuel consumption for passenger vehicles in three different driving styles. By using the results of this study, performance of different data analytics tools can be compared based on different scenario types.

The results of this work will help improving of vehicle design phase in term of vehicle's environmental certification. On the other hand, embedding the results of this paper into mathematical models will improve vehicle routing problems and simulation-based evaluations.

The rest of this study is organized as follows: **Section 2** gives an overview of related works. **Section 3** explains model factors, data characteristics and descriptions. **Section 4** describes methodologies and the adopted algorithms. **Section 5** covers empirical analysis and shows how to select the best-fitted algorithms, as well as offering a practical challenge to the methodology, and performance analysis. Then, an interpretation of the results is provided. **Section 6** challenges some benchmarks by comparing their methods' performances with that of the proposed approach. In **Sections 7** and **8**, one can find a discussion of the work's implications and the conclusion of the paper.

LITERATURE REVIEW

In 2010, there were about 1 billion cars and trucks on the road worldwide. For 2030, it is expected that number will reach about 1.7 billion, with the strongest growth taking place in Asia and the Middle East according to EU

Vehicle Market Statistics [6]. Analysis of "Big Data" is the greatest asset in the quest to reduce fuel expenses and will enable companies to exploit opportunities to gain a competitive advantage. Accurate data, analyzed correctly, and leveraged with clean diesel technology, will create a significant cost savings for fleet operations and will provide the added benefit of lowering its carbon footprint [7]. Zhang et al. (2014), aimed to evaluate real-world fuel consumption and CO_2 emission factors for Light-Duty Passenger Vehicles (LDPVs) in China, discuss the impacts from on-road driving conditions and further stress the importance of reducing fuel consumption through effective traffic measures. The results can improve policy-makers' understanding of the role of real-world traffic conditions in influencing fuel consumption and CO_2 emissions of LDPVs.

For the same aim, Ehsani, Ahmadi, & Fadai (2016) investigated the effectiveness of new parameters such as temperature, driving style, asphalt efficiency and fuel efficiency by proposing a model based on top-down mechanical model which has five parts including the energy consumed due to loss of gravity, acceleration, rolling resistance, aerodynamic resistance and cornering losses.

Wang, Zhang, & Ouyang (2015), assessed fuel reduction for three types of electric vehicles by analyzing driving patterns and comparing them with conventional vehicles. They have found that low speeds, long idle times, severe speed changes and short driving ranges are the main driving patterns in which electric vehicles are more advantageous than conventional vehicles in the aspect of fuel reduction. They have also found that smaller batteries with 30-50 km Charging Depleting range would meet the drivers' expectations and help to minimize the costs to the vehicle.

Li, Wang, & Song, (2016) has investigated fuel consumption in a car-following process to optimize it for smart hybrid electric vehicles. They have proposed a novel optimization algorithm which is based on Model Predictive Control (MPC), and carried out a simulation in MATLAB to evaluate the results. It is found that the proposed method is more effective as much as 5% in reducing fuel consumption than energy management strategy and adaptive cruise control together.

Constantinescu, Marinoiu, & Vladoiu (2010) have investigated driving patterns by using fleet management data to increase traffic safety. They have collected the driving parameters by an in-house developed GPS-based device (Gipix) from urban traffic and analyzed by using Hierarchical Cluster Analysis and Principal Component Analysis and after gathering them from data server. They have found that driving environment, traffic context, driver's particular condition, and driver's characteristics play a role in driving behavior.

X. Wang et al., (2015) have also studied on finding driving patterns by creating a statistical pattern recognition technique using Bayesian Probability and Kernel Density Estimation. In this study, uncertainty of driving styles has been taken into account and drivers have been categorized. Benefits of the new model have shown by using a cross validation technique. According to the results the model can improve the recognition both for aggressive and normal drivers and has a higher stability than fuzzy logic algorithm.

Real-time fuel usage data of a single vehicle on a specified road also collected and analyzed by using different emission models to show how strong they are in estimation of carbon dioxide emission.

DATA CHARACTERISTICS OF THE MODEL

In this study, car fuel consumption and emissions data of used passenger cars which are registered in 2015 from Vehicle Certification Agency [14] and vehicle specification data of Mercedes Benz [15] have been used. Initial dataset used in this study consists of 2667 cases and 27 variables. However, all data were in one column in Excel. That is why they are separated into the columns, such as; Manufacturer, Model, Description, Transmission, Engine Capacity, Fuel Type, Electric energy consumption Miles/kWh, wh/km, Maximum range (Km), Maximum range (Miles), Metric Urban (Cold), Metric Extra-Urban, Metric Combined, Imperial Urban (Cold), Imperial Extra-Urban, Imperial Combined, CO_2 g/km, Fuel Cost 12000 Miles, Electricity cost, Total cost/12000 miles, Euro Standard, Noise Level [dB(A)], Emissions CO [mg/km], THC Emissions [mg/km], Emissions NOx [mg/km], THC + NOx Emissions [mg/km], Particulates [No.] [mg/km], by using tools of Excel.

PROPOSED METHODOLOGY AND ALGORITHMS

1. Design Phase

Modeling is the flashy part of data mining, where sophisticated analysis methods are used to extract information from the data. This phase involves selecting modeling techniques, generating test designs, and building and assessing models. For assessment, some approaches based on multi criteria decision making and data envelopment analysis have already been developed [16].

In this study, raw data has been gathered from Vehicle Certification Agency and website of Mercedes Benz. Steps of application has been designed for this study in the as shown in **Figure 1** and has been followed.

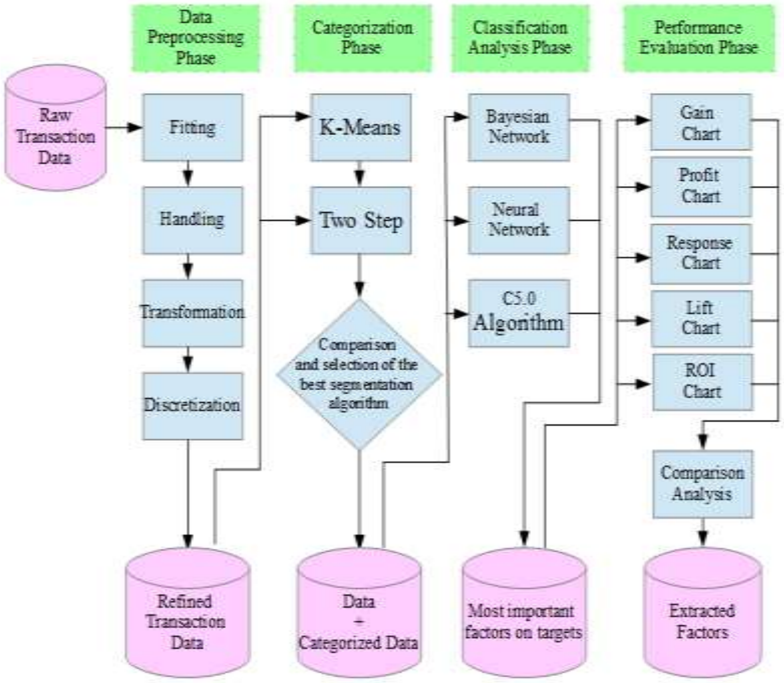


Figure 1: Proposed methodology and operation steps flowchart

Euro 6 data was not suitable for an analysis and needed to be processed. For this, data preprocessing steps of a proposed methodology has taken into account [17]. By using technical specification data, new variables and other missing values have been filled. Outliers and inaccurate cases have deleted in compliance with their statistics. Dataset has been transformed into a structure which can be used in following analysis, while continuous variables have been splitted into smaller groups of dataset. Refined dataset has been kept and categorized by K-means and Two-step algorithms. Results of categorization have compared to each other and the best segmentation algorithm has chosen. Categorized data has been saved with initial data and has been used in classification analysis.

Bayesian network, neural network and C5.0 algorithm were used in classification analysis phase. These have found the most important factors effective on desired targets such as; fuel efficiency, emissions and noise level. Most important factors found by each technic has been kept to visualize by several charts, such as; gain charts, profit charts, response charts, lift charts and return on investment charts. In performance evaluation phase, results of those graphics were compared and analyzed, then extracted factors have been kept.

1.1 Learning Algorithms

1.1.1 Clustering (Segmentation)

Clustering divides a database into different groups [18]. The goal of clustering is to find groups that are very different from each other, and whose members are very similar to each other. Often it is necessary to modify the clustering by excluding variables that have been employed to group instances, because upon examination the user identifies them as irrelevant or not meaningful. After finding clusters that reasonably segment database, these clusters may then be used to classify new data. Mooi, Erik; Sarstedt (2011) have explained partition algorithms and steps of clustering analysis.

In this study, K-means and Two-step clustering algorithms have been used.

Two-Step

The name two-step clustering is already an indication that the algorithm is based on a two-stage approach: In the first stage, the algorithm undertakes a procedure that is very similar to the k-means algorithm. Based on these results, the two-step procedure conducts a modified hierarchical agglomerative clustering procedure that combines the objects sequentially to form homogenous clusters. This is done by building a so-called cluster feature tree whose "leaves" represent distinct objects in the dataset. The procedure can handle categorical and continuous variables simultaneously and offers the user the flexibility to specify the cluster numbers as well as the maximum number of clusters, or to allow the technique to automatically choose the number of clusters on the basis of statistical evaluation criteria. Furthermore, the procedure indicates each variable's importance for the construction of a specific cluster. These desirable features make the somewhat less popular two-step clustering a viable alternative to the traditional methods [19].

K-means

The simplest and most commonly used algorithm, employing a squared error criterion is the K-means algorithm [20]. This algorithm partitions the data into K clusters$(C_1 C_2, \ldots, C_K)$, represented by their centers or means. The center of each cluster is calculated as the mean of all the instances belonging to that cluster.

The algorithm starts with an initial set of cluster centers, chosen at random or according to some heuristic procedure. In each iteration, each instance is assigned to its nearest cluster center according to the Euclidean distance between the two. Then the cluster centers are re-calculated. The center of each cluster is calculated as the mean of all the instances belonging to that cluster:

$$\mu_k = \frac{1}{N_k} \sum_{q=1}^{N_k} x_q$$

where N_k is the number of instances belonging to cluster k and μ_k is the mean of the cluster k.

Linear complexity is one of the reasons for the popularity of the K-means algorithms. Even if the number of instances is substantially large (which often is the case nowadays), this algorithm is computationally attractive. Thus, the K-means algorithm has an advantage in comparison to other clustering.

At clustering step, output and quality of the clusters in K-means algorithm largely depend on selecting the proper number of clusters. For selecting the optimal number of clusters, there are many indicators such as Davies–Bouldin Index, the Silhouette Width or Dunn Index. In this work we are using Dunn Index. The Dunn's index measures compactness (Maximum distance in between data points of clusters) and clusters separation (minimum distance between clusters). This measurement serves as a measure to find the right number of clusters in a data set, where the maximum value of the index represents the right partitioning given the index [21]. The aim of all these indexes is to have meaningful clusters where the data objects within the same cluster are similar to one another and dissimilar to the objects in other clusters. In this research, we evaluate the optimum number of clusters based on Dunn Index in order to use in K-means algorithm.

1.1.2 Classification

After analysis, to have a clear view of results and a general outcome are needed and having sets would be explicit to explain. This can be done on SPSS by grouping due to range and desired number of sets, but classification algorithms give more meaningful categories. This is why the classification algorithms, such as; neural network, Bayesian network and C5.0 have been used in this study.

Bayesian Network Model

A Bayesian network is a graphical model that displays variables (often referred to as nodes) in a dataset and the probabilistic, or conditional, independencies between them. Bayes networks as implemented SPSS are used to predict categorical outcomes [22]. A Bayesian network comprises two main components [23]. The first is an acyclic oriented graph in which the nodes correspond to the predictive variables and the arcs indicate relationships of stochastic dependence. In particular, it is assumed that the variable X_j associated with node aj in the network is dependent on the variables associated with the predecessor nodes of a_j, and conditionally independent of the variables associated with the nodes that are not directly reachable from a_j. The second component consists of a table of conditional probabilities assigned for each variable. In particular, the table associated with the variable X_j indicates the conditional distribution of $P(X_j\,|C_j)$, where C_j represents the set of explanatory variables associated with the predecessor nodes of node aj in the network and is estimated based on the relative frequencies in the dataset.

Neural Network Model

Neural networks are intended to simulate the behavior of biological systems composed of neurons. Since the 1950s, when the simplest models were proposed, neural networks have been used for predictive purposes, not only for classification but also for regression of continuous target attributes [23]. A neural network is an oriented graph consisting of nodes, which in the biological analogy represent neurons, connected by arcs, which correspond to dendrites and synapses. Each arc is associated with a weight, while at each node an activation function is defined which is applied to the values received as input by the node along the incoming arcs, adjusted by the weights of the arcs. The training stage is performed by analyzing in sequence the observations contained in the training set one after the other and by modifying at each iteration the weights associated with the arcs.

2. Analysis Phase

Official fuel consumption test procedures have been in use since the 1970s. EU Directive 80/1268/EEC as amended or, for Euro 5 vehicles onwards, Regulation 692/2008 describe the tests which all new cars on sale after 1 January 2001 are required to take [24]. The current test for conventional internal combustion engine vehicles has three parts. These are an urban, and an extra-urban, and combined urban cycle. The test cycle is the same as that used to determine the official exhaust air quality pollutant emission classification for the model of vehicle in question. The cars tested have to be 'run-in' so they must have been driven for at least 1,800 miles (3,000 kilometers) before testing.

For evaluating of emission levels a different test environment is needed. For this reason, much of the testing revolved around the use of the New European Driving Cycle (NEDC), which is the current emission test for vehicles in type approval and therefore the cycle that such strategies would be designed to recognize.

The test program is constructed around variations of this cycle with testing being undertaken in emissions laboratories. Further tests were conducted on public roads to establish the emissions performance of the vehicles in typical real-world use conditions. The United Nations is currently developing a new vehicle test procedure, the Worldwide Harmonized Light Vehicles Test Procedure (WLTP), which is close to being finalized and will address some of the known issues with current vehicle testing. It is expected that the WLTP will be implemented in the EU by 2016/2017.

3. Performances Evaluation Phase

Evaluation Node of Clementine has been used to evaluate the performance of the three classification methods for each type of fuel consumption, level of emissions and noise level. The Evaluation node offers an easy way to evaluate and compare predictive models to choose the best model for an application. Evaluation charts show how models perform in predicting particular outcomes. They work by sorting records based on the predicted value and confidence of the prediction, splitting the records into groups of equal size (quantiles), and then plotting the value of the business criterion for each quantile, from highest to lowest. Multiple models are shown as separate lines in the plot.

Outcomes are handled by defining a specific value or range of values as a hit. Hits usually indicate success of some sort (such as a sale to a customer) or an event of interest (such as a specific medical diagnosis). Hit criteria can be defined on the Options tab of the dialog box, or the default hit criteria can be used as follows:

- Flag output fields are straightforward; hits correspond to true values.
- For Set output fields, the first value in the set defines a hit.
- For Range output fields, hits equal values greater than the midpoint of the field's range.

There are five types of evaluation charts, each of which emphasizes a different evaluation criterion.
Gains Charts: Gains are defined as the proportion of total hits that occurs in each quantile. Gains are computed as (number of hits in quantile / total number of hits) × 100%.

Lift Charts: Lift compares the percentage of records in each quantile that are hits with the overall percentage of hits in the training data. It is computed as (hits in quantile / records in quantile) / (total hits / total records).

Response Charts: Response is simply the percentage of records in the quantile that are hits. Response is computed as (hits in quantile / records in quantile) × 100%.

Profit Charts: Profit equals the revenue for each record minus the cost for the record. Profits for a quantile are simply the sum of profits for all records in the quantile. Revenues are assumed to apply only to hits, but costs apply to all records. Profits and costs can be fixed or can be defined by fields in the data. Profits are computed as (sum of revenue for records in quantile − sum of costs for records in quantile).

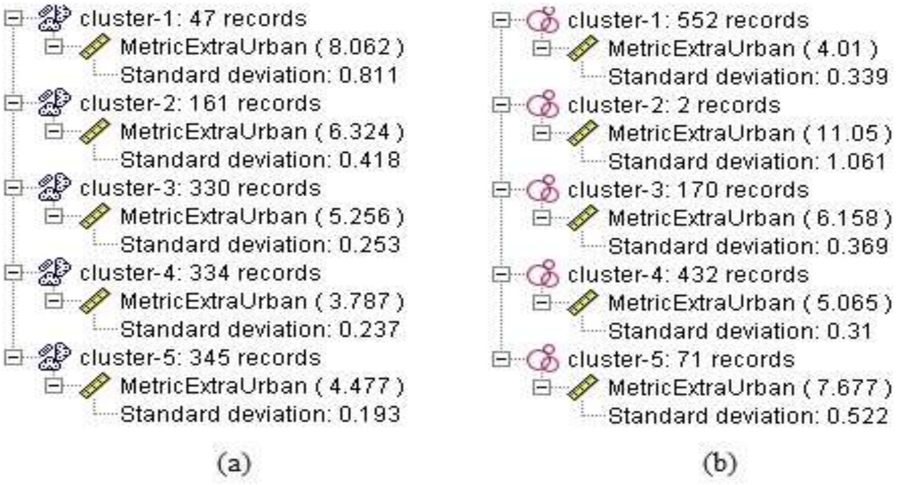
ROI Charts: ROI (return on investment) is similar to profit in that it involves defining revenues and costs. ROI compares profits to costs for the quantile. ROI is computed as (profits for quantile / costs for quantile) × 100%.

EMPIRICAL ANALYSIS AND RESULTS

Metric Extra Urban

The extra urban cycle is a cycle that is intended to represent the use of the vehicle on roads that are external to the urban environment [24]. The cycle is conducted immediately following the urban cycle and consists of roughly half steady speed driving with the remainder being accelerations, decelerations, and some idling. The maximum speed is 75 mph (120 km/h). The average speed is 39 mph (63 km/h) and the distance covered is 4.3 miles (7 km). For instance, figure 2 shows that 5 clusters (sets) for metric extra urban (cold) (l/100 km) have been obtained by using Two Step algorithm as many as K-means algorithm has found, although the centers of the sets found by each technic are different.

Most of the metric extra urban values appeared in cluster-5 with 345 cases in Two-Step, while they belonged to cluster-1 with 552 cases in K-means. With Two step algorithm, each cluster has had the central values of 8.06 l, 6.32 l, 5.26 l, 3.79 l and 4.48 l, while clusters of K-means have had 4.01 l, 11.05 l, 6.16 l, 5.06 l, and 7.68 l, respectively.



```
cluster-1: 47 records
    MetricExtraUrban ( 8.062 )
        Standard deviation: 0.811
cluster-2: 161 records
    MetricExtraUrban ( 6.324 )
        Standard deviation: 0.418
cluster-3: 330 records
    MetricExtraUrban ( 5.256 )
        Standard deviation: 0.253
cluster-4: 334 records
    MetricExtraUrban ( 3.787 )
        Standard deviation: 0.237
cluster-5: 345 records
    MetricExtraUrban ( 4.477 )
        Standard deviation: 0.193
```

(a)

```
cluster-1: 552 records
    MetricExtraUrban ( 4.01 )
        Standard deviation: 0.339
cluster-2: 2 records
    MetricExtraUrban ( 11.05 )
        Standard deviation: 1.061
cluster-3: 170 records
    MetricExtraUrban ( 6.158 )
        Standard deviation: 0.369
cluster-4: 432 records
    MetricExtraUrban ( 5.065 )
        Standard deviation: 0.31
cluster-5: 71 records
    MetricExtraUrban ( 7.677 )
        Standard deviation: 0.522
```

(b)

Figure 2: Clusters for metric extra urban by (a) Two-Step and (b) K-Means algorithms

Metric Urban (Cold)

The urban test cycle is carried out in a laboratory at an ambient temperature of 20°C to 30°C on a rolling road from a cold start where the engine has not run for several hours [24]. The cycle consists of a series of accelerations, steady speeds, decelerations and idling. The maximum speed is 31 mph (50 km/h). The average speed 12 mph (19 km/h) and the distance covered is 2.5 miles (4 km). Analysis shows that 5 clusters (sets) for metric urban (cold) (l/100 km) have been obtained by using Two Step algorithm as well as K-means algorithm.

Metric Combined

The combined (l/100km) is for the urban (cold) and the extra-urban cycle together. It is therefore an average of the two other parts of the fuel consumption test, urban and extra-urban cycles, weighted by the distance covered in each part [24]. 3 clusters (sets) for metric combined have been obtained by using Two Step algorithm while K-means algorithm has found 4 clusters. Analysis shows that 4 clusters (sets) for carbon monoxide have been obtained by using Two Step algorithm while K-means algorithm has found 5 clusters.

CO Emission

Carbon monoxide reduces the bloods oxygen carrying capacity which can reduce availability of oxygen to key organs. This is why amount of CO emission (gram) produced per 100 kilometers is evaluated for passenger cars [14].

CO_2 Emission

When petrol or diesel is burnt for energy the main by-products are water vapour and carbon dioxide (CO_2). Carbon dioxide is the most important of the greenhouse gases which are contributing to climate change. In 6000 miles a car will produce roughly its own weight in CO_2. Cars registered on or after 1 March 2001 are taxed according to the amount of CO_2 emitted and the type of fuel used [14]. Analysis resulted that importance of each cluster with two step algorithm have been ordered as cluster-3, cluster-4, cluster-2, and cluster-1, respectively. With K-means, order of importance of the sets has been cluster-2, cluster-3, cluster-1, cluster-4 and cluster-5, respectively.

THC Emission

Hydrocarbons contribute to ozone formation. Some kinds of HCs can also be carcinogens and they are also indirect greenhouse gases [14]. Total hydrocarbon emission (gram) has been measured per kilometer. Analysis shows that 4 clusters (sets) for hydrocarbon emission have been obtained by using Two Step algorithm while K-means algorithm has found 5 clusters.

NO_x Emission

Nitrogen oxides (NO_X) can have serious adverse health effects. In particular, nitrogen dioxide is the main precursor to ground-level ozone, which causes significant respiratory problems that can lead to premature death [25]. The NO_x limit in the current Euro 6 standard, introduced in 2014, is over 90% lower. However emerging evidence has shown that the main reason that air quality standards have not improved, despite these increasingly stringent regulatory standards, is that vehicles in the ‘real world’ are not performing as well as when tested in the laboratory according to The UK Department for Transport (2016). Analysis resulted that, 4 clusters (sets) for NO_x emission have been obtained by using Two Step algorithm while K-means algorithm has found 5 clusters.

Particulate Emissions

Fine particles (especially those smaller than 10 microns) enter the lungs and are associated with a range of respiratory and cardiovascular disorders and asthma[14]. Amount of particulate emissions is measured per 100 kilometers. As analysis results, 4 clusters (sets) for hydrocarbon emission have been obtained by using Two Step algorithm while obtaining 5 clusters by K-means.

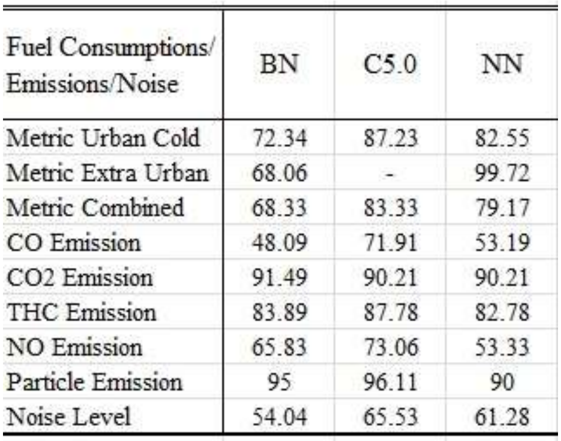
Noise Level

All passenger cars, from small hatchbacks to sports cars, large luxury saloons and 9-seat people carriers fall into the M1 regulatory category, with varying requirements for sound levels [27]. To improve the situation, both the industry and the Commission has proposed a passenger car classification system which is dependent on the power-to-mass ratio of the different vehicle families. The proposed classification divides the existing M1 category into 3 segments, with 99% of vehicles falling into the green ‘normal’ category. Analysis shows that 5 clusters (sets) for noise level have been obtained by using Two Step algorithm as well as obtaining 5 clusters by K-means.

Factor Evaluation

After analysis, successes of Bayesian network, neural network and C5.0 algorithm in prediction of fuel consumption, emission and noise level metrics have been evaluated by comparing with each other (**Table 1**).

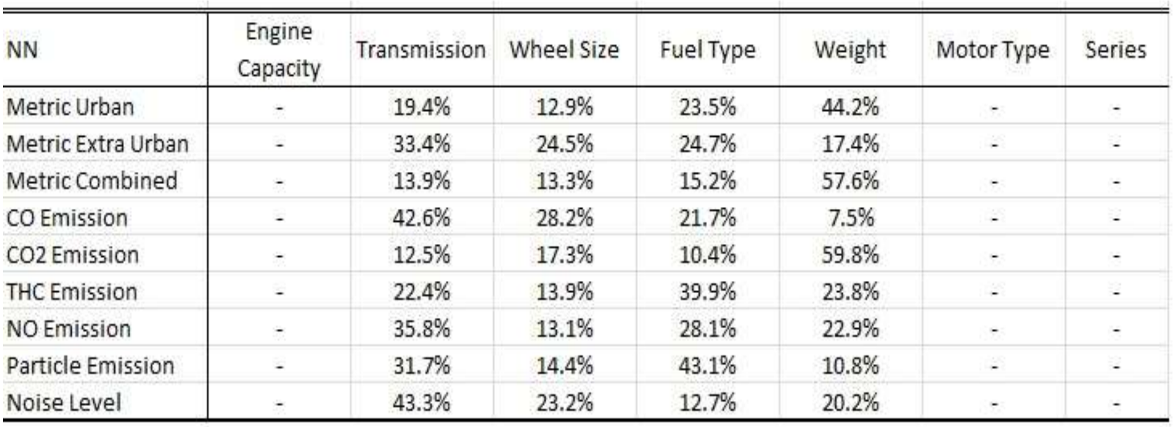
C5.0 algorithm has shown a better prediction for seven metrics (metric urban cold, metric combined, CO emission, THC emission, NO emission, particle emission and noise level) however it get no feasible solution in predicting parameters for metric extra urban. Instead, neural network has predicted parameters for metric extra urban with nearly 100% success. For CO_2 emission, Bayesian network has been the best technic to use. Average successes have been 72%, 82% and 77% for Bayesian network, C5.0 and neural network, respectively.



Fuel Consumptions/ Emissions/Noise	BN	C5.0	NN
Metric Urban Cold	72.34	87.23	82.55
Metric Extra Urban	68.06	-	99.72
Metric Combined	68.33	83.33	79.17
CO Emission	48.09	71.91	53.19
CO2 Emission	91.49	90.21	90.21
THC Emission	83.89	87.78	82.78
NO Emission	65.83	73.06	53.33
Particle Emission	95	96.11	90
Noise Level	54.04	65.53	61.28

Table 1: Prediction success of classification algorithms

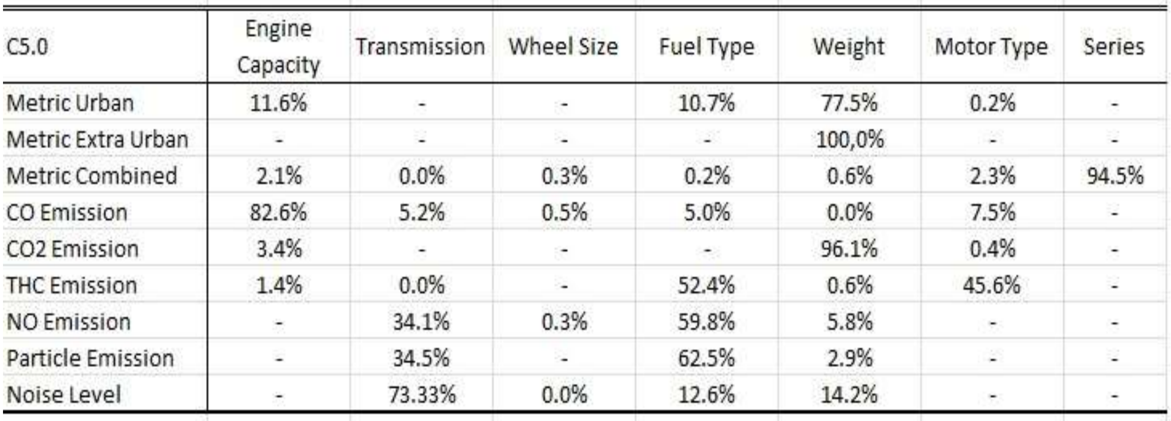
Comparisons have done also for parameters which effect fuel consumptions, emissions and noise level. In prediction of fuel consumption by neural network (**Table 2**), weight, transmission and fuel type of passenger cars have been the most important indicators, while transmission, fuel type, weight have been the most important ones for emissions. Transmission, wheel size and weight, respectively, effect noise level of vehicles at most.



NN	Engine Capacity	Transmission	Wheel Size	Fuel Type	Weight	Motor Type	Series
Metric Urban	-	19.4%	12.9%	23.5%	44.2%	-	-
Metric Extra Urban	-	33.4%	24.5%	24.7%	17.4%	-	-
Metric Combined	-	13.9%	13.3%	15.2%	57.6%	-	-
CO Emission	-	42.6%	28.2%	21.7%	7.5%	-	-
CO2 Emission	-	12.5%	17.3%	10.4%	59.8%	-	-
THC Emission	-	22.4%	13.9%	39.9%	23.8%	-	-
NO Emission	-	35.8%	13.1%	28.1%	22.9%	-	-
Particle Emission	-	31.7%	14.4%	43.1%	10.8%	-	-
Noise Level	-	43.3%	23.2%	12.7%	20.2%	-	-

Table 2: Weights of parameters found by neural network

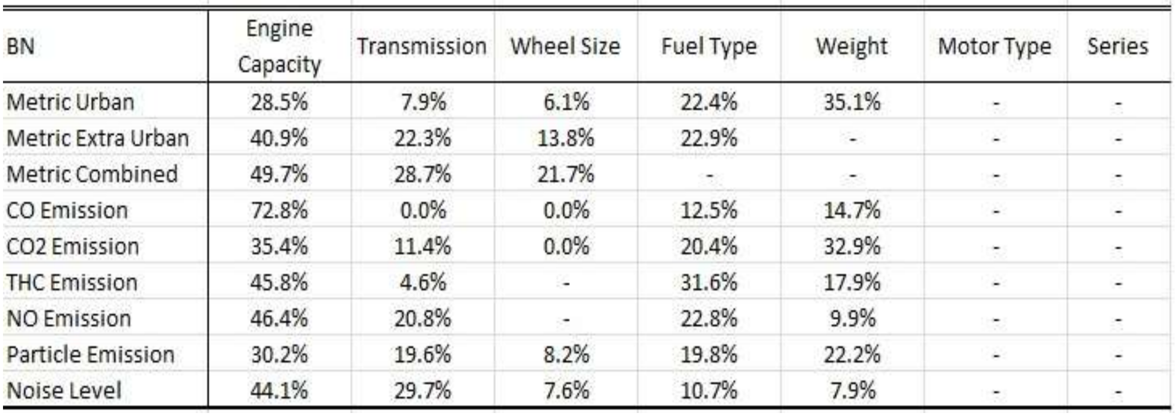
In prediction of fuel consumption by C5.0 algorithm (**Table 3**), weight has effected all metrics, especially fuel consumption metrics. Series of models of cars were very important for metric combined. For emissions, fuel type has been the most effective parameter and engine capacity, weight, transmission, motor type have followed, respectively. Wheel sizes of cars are not significantly effective on both fuel consumptions and emissions due to its low average degree (0.4%) as well as for noise level. Similar to results of neural network, transmission effects noise level more than other parameters.



C5.0	Engine Capacity	Transmission	Wheel Size	Fuel Type	Weight	Motor Type	Series
Metric Urban	11.6%	-	-	10.7%	77.5%	0.2%	-
Metric Extra Urban	-	-	-	-	100,0%	-	-
Metric Combined	2.1%	0.0%	0.3%	0.2%	0.6%	2.3%	94.5%
CO Emission	82.6%	5.2%	0.5%	5.0%	0.0%	7.5%	-
CO2 Emission	3.4%	-	-	-	96.1%	0.4%	-
THC Emission	1.4%	0.0%	-	52.4%	0.6%	45.6%	-
NO Emission	-	34.1%	0.3%	59.8%	5.8%	-	-
Particle Emission	-	34.5%	-	62.5%	2.9%	-	-
Noise Level	-	73.33%	0.0%	12.6%	14.2%	-	-

Table 3. Weights of parameters found by C5.0

Bayesian network have found that engine capacity and transmission effect all metrics (**Table 4**). To predict fuel consumption, engine capacity (39,7%) and weight (35,1%) have been the most important parameters. Engine capacity also effects noise level more than other parameters. For emissions, engine capacity (46,1%), fuel type (21,4%), weight (15,5%), transmission (11,3%) have been the most important parameters, respectively.



BN	Engine Capacity	Transmission	Wheel Size	Fuel Type	Weight	Motor Type	Series
Metric Urban	28.5%	7.9%	6.1%	22.4%	35.1%	-	-
Metric Extra Urban	40.9%	22.3%	13.8%	22.9%	-	-	-
Metric Combined	49.7%	28.7%	21.7%	-	-	-	-
CO Emission	72.8%	0.0%	0.0%	12.5%	14.7%	-	-
CO2 Emission	35.4%	11.4%	0.0%	20.4%	32.9%	-	-
THC Emission	45.8%	4.6%	-	31.6%	17.9%	-	-
NO Emission	46.4%	20.8%	-	22.8%	9.9%	-	-
Particle Emission	30.2%	19.6%	8.2%	19.8%	22.2%	-	-
Noise Level	44.1%	29.7%	7.6%	10.7%	7.9%	-	-

Table 4: Weights of parameters found by Bayesian network

Benchmarking

To improve the quality of the work, we compared our proposed methodological techniques with similar works that have been conducted lately. Different algorithms from similar works were applied to our data set. The results of the proposed methodology and the results of the benchmarks are summarized in **Table 5**. Clearly the gained results from proposed algorithms; apart from NOx and Noise level, outstandingly are better than other benchmarks.

DISCUSSION

What we have done in this paper was to pursue a strategy to understand how the application of big data mining in the automotive industry does potentially works, targeting fuel consumptions and emission levels of passenger cars from different firms. With awareness of the fact that, big data caused an explosion in the use of more extensive data mining techniques, partially because the size of the information is much larger and the information tends to be more varied and extensive in its very nature and content with large datasets, it is no longer enough to get relatively simple and straightforward statistics out of the system.

		Urban Metric Cold	Urban Metric Extra	Urban Metric Combined	CO	CO2	NOx	Particulates	THC	Noise
Work	Method	Test	Test	Test	Test	Test	Test	Test	Test	Test
Proposed	NN	82.5	99.7*	79.1	53.1	90.2	53.3	90	82.7	61.2
Proposed	C5.0	87.2*	-	83.3*	71.9*	90.2	73.0	96.1*	87.7*	65.5
Proposed	Bayes Net.	72.3	68.6	68.3	48.0	91.4*	65.8	95	83.8	54.0
Haifeng, Xu-hui,& Xin-feng, 2015	SVM	63.8	73.2	61.7	53.6	74.2	80.1*	74.6	70.3	70.3
CGI GROUP INC., 2014)	CHAID	82.9	72.5	65.4	60.1	64.7	72.2	74.5	70.8	80.2*
(Viswanathan, 2013	C&R Tree	61.2	49.9	53.2	54.2	68.4	72.4	63.4	77.2	58.5

Table 5: Summarized benchmarking results (*the best scenario)

Considering the limitations in data acquisition phase of this paper, we used a relatively normal size of an extracted dataset aiming to test different technics and approaches of data mining in term of verifying a proposed methodology. Considering this aspect that fundamentally, data mining is about processing data and identifying patterns and trends in that information so that it can be decided or judged. On the other hand, data mining is not all about the tools or database systems and simple tools, including creating and writing our own, or using off the shelf software packages. Complex data mining benefits from the past experience and algorithms defined with existing software and packages, with certain tools gaining a greater affinity or reputation with different techniques. In context of design and production of automobiles, it may have assumed that data mining principles have been around for many years, but, with the advent of big data, it is even more prevalent and consequently it resulted to the exact concept of connected cars. Results of this paper show the abilities of data mining tools in design and extension phases of automotive industry. Furthermore, it is depicting a road map of connection between collected data from connected cars and the cluster and large-scale data processing are able to allow data mining to collate and report on groups and correlations of data that are more complicated. Progressing data mining tools will certainly provide an entirely new range of tools and systems available, including combined data storage and processing systems for automotive industry.

CONCLUSION AND FUTURE WORKS

This paper is an empirical analysis of fuel consumptions and emission levels of passenger cars in context of application potential of big data mining. In this way, the dataset which has been resulted from combining Euro 6 and technical specification data of Mercedes Benz, BMW and Audi automobiles (passenger cars) have been used. Results of descriptive statistics and information about data have collected by different data mining tools. Usages have been compared and parameters have been derived by segmentation algorithms aiming better results by categorizing of variables for upcoming analysis. All data mining applications have been designed and carried out by following the steps of CRISP DM.

The results obtained from adapting of various machine learning tools and applications has shown that the most important parameters differ based on applied technic. C5.0 algorithm has given a better prediction for most of the cases.

Additionally, the importance of each parameter has been evaluated to predict its contribution on fuel consumption, on emission level and noise level by data analytics technics. C5.0 algorithm has found the weight factor more effective on fuel consumption, fuel type on emissions and transmission on noise level. Engine capacity and weight have been the most important parameters for Bayesian network. Engine capacity also effects noise level more than other parameters. Results of neural network have shown that weight, transmission and fuel type of passenger cars have been the most important indicators for fuel consumption and for emissions. Transmission, wheel size and weight, respectively, effect noise level of vehicles at most. A benchmarking impended in order to challenge other technics and methodologies in recent literature. Benchmarking results have been verified this paper' approach of being on right track in terms of methodology and adopted technics.

Different data analytics tools can be compared based on different scenario types by using the results of this study. The implications of this work can be taken account in design phase of automobiles' urban tests. Also, the results of this work will help improving of vehicle design phase in term of vehicle's environmental certification. On the other hand, embedding the results of this paper into mathematical models will improve vehicle routing problems and simulation-based evaluations.

REFERENCES

[1] T. King, "The automotive industry," *Eur. Voice*, vol. 13 (16), no. 3, p. p16, 2007.

[2] M. Johanson, S. Belenki, J. Jalminger, M. Fant, and M. Gjertz, "Big automotive data: Leveraging large volumes of data for knowledge-driven product development," *Proc. - 2014 IEEE Int. Conf. Big Data, IEEE Big Data 2014*, pp. 736–741, 2015.

[3] A. Hiziroglu, "Soft computing applications in customer segmentation: State-of-art review and critique," *Expert Syst. Appl.*, vol. 40, no. 16, pp. 6491–6507, 2013.

[4] "Does Your Company Have a Data Science Strategy?," *Harvard Business Review*, 2016. [Online]. Available: https://hbr.org/sponsored/2016/08/does-your-company-have-a-data-science-strategy.

[5] C. Lynette, E. Christopher, B. Anup, and H. John, "Factor of Two: Halving the Fuel Consumption of New U.S. Automobiles by 2035," no. October, pp. 1–51, 2007.

[6] ICCT Europe, *EUROPEAN VEHICLE MARKET STATISTICS Pocketbook 2015/16*. 2013.

[7] Fleet Owner, "Leveraging 'Big Data' to Maximize Fuel Economy," no. October, pp. 1–3, 2009.

[8] S. Zhang, Y. Wu, H. Liu, R. Huang, P. Un, Y. Zhou, L. Fu, and J. Hao, "Real-world fuel consumption and CO2 (carbon dioxide) emissions by driving conditions for light-duty passenger vehicles in China," *Energy*, vol. 69, pp. 247–257, 2014.

[9] M. Ehsani, A. Ahmadi, and D. Fadai, "Modeling of vehicle fuel consumption and carbon dioxide emission in road transport," *Renew. Sustain. Energy Rev.*, vol. 53, pp. 1638–1648, 2016.

[10] H. Wang, X. Zhang, and M. Ouyang, "Energy consumption of electric vehicles based on real-world driving patterns : A case study of Beijing q," *Appl. Energy*, vol. 157, pp. 710–719, 2015.

[11] L. Li, X. Wang, and J. Song, "Fuel consumption optimization for smart hybrid electric vehicle during a car-following process," 2016.

[12] Z. Constantinescu, C. Marinoiu, and M. Vladoiu, "Driving Style Analysis Using Data Mining

Techniques," *Int. J. Comput. Commun. Control*, vol. 5, no. 5, pp. 654–663, 2010.

[13] X. Wang, J. C. Chow, S. D. Kohl, K. E. Percy, A. H. Legge, and J. G. Watson, "Real-world emission factors for Caterpillar 797B heavy haulers during mining operations," *Particuology*, vol. 28, pp. 22–30, 2015.

[14] I. Band, "Car fuel data, CO2 and vehicle tax," *United Kingdom Vehicle Certification Agency*, 2016. [Online]. Available: http://carfueldata.direct.gov.uk/.

[15] "Technical Specification," *Mercedes-Benz UK Ltd.*, 2004. [Online]. Available: http://www.mercedes-benz.co.uk/content/unitedkingdom/mpc/mpc_unitedkingdom_website/en/home_mpc/passengercars/home/new_cars/models/e-class/_s212/advice_sales/technical-spec.html#.

[16] G. Nakhaeizadeh and A. Schnabl, "Development of Multi-Criteria Metrics for Evaluation of Data Mining Algorithms," *Int. Conf. Knowl. Discov. Database Data Min.*, pp. 37–42, 1997.

[17] P. A. Sarvari, A. Ustundag, and H. Takci, "Performance evaluation of different customer segmentation approaches based on RFM and demographics analysis," *Kybernetes*, vol. 45, no. 7, pp. 1129–1157, 2016.

[18] H. A. Edelstein, *Introduction to Data Mining and Knowledge Discovery*, vol. 2. 1999.

[19] M. Mooi, Erik; Sarstedt, *A Concise Guide to Market Research The Process, Data, and Methods Using IBM SPSS Statistics Introduction to Market Research*, vol. 68, no. 1. Heidelberg, 2011.

[20] L. Rokach and O. Maimon, "Chapter 15— Clustering methods," *Data Min. Knowl. Discov. Handb.*, p. 32, 2010.

[21] M. K. Pakhira, S. Bandyopadhyay, and U. Maulik, "Validity index for crisp and fuzzy clusters," *Pattern Recognit.*, vol. 37, no. 3, pp. 487–501, 2004.

[22] F. P. Di Marco and R. T. Hanlon, "Chapter 11.pdf," *Book*, vol. 103. pp. 29–38, 2003.

[23] D. R. Tobergte and S. Curtis, *Business Intelligence*, vol. 53, no. 9. 2013.

[24] V. V. C. A. Offices, H. Page, H. Pagewhat, A. This, S. Vcapublicationsmedia, A. Home, F. Consumption, and E. Union, "The fuel consumption testing scheme," 2016. .

[25] V. Franco, F. P. Sánchez, J. German, and P. Mock, "FACT SHEET : EUROPE REAL-WORLD EMISSIONS FROM MODERN DIESEL CARS," *Int. Counc. CLEAN Transp.*, vol. 45, no. X, pp. 1–4, 2014.

[26] The UK Department for Transport, "Vehicle Emissions Testing Programme - Moving Britain Ahead," no. April, 2016.

[27] ACEA, *Vehicle Noise - Setting the right sound levels*. 2012.

[28] CGI GROUP INC., "Modeling the Relation Between Driving Behavior and Fuel Consumption," pp. 1–10, 2014.