WORKING PAPER
ITS-WP-99-10

Suitability of Fuel Efficiency as a Criterion in Passenger Vehicle Classification: An Investigation of the Classification Capability of Decision Tree Approach

by

Tu T. Ton
& Baojin Wang

May, 1999

ISSN 1440-3501

Established and supported under the Australian Research Council's Key Centre Program.

INSTITUTE OF TRANSPORT STUDIES
The Australian Key Centre
in Transport Management

The University of Sydney
and Monash University

Introduction

Studies of the demand for automobiles where an emphasis is on the class of vehicle typically use a number of physical and performance attributes to group vehicles, treating them as if they are homogeneous in respect of a particular application. A most common application is the prediction of energy consumed and its conversion into greenhouse gas emissions. Since fuel efficiency, a major component of the calculation of CO_2 emissions, has not been used as a classification criterion, it is unclear as to how suitable the existing vehicle classes are in studying the environmental impact of policies designed to impact on the demand for automobiles by class.

In addressing this issue, this paper employs the classification and regression trees (CART) approach to identify the suitability of the existing vehicle classification scheme for environmental and energy-based applications. The paper is structured around five sections. Next section reviews current passenger vehicle classification scheme and the associated issues. Section three describes the background and motivations for using CART to build vehicle classification models. The development and evaluation of vehicle classification models with and without fuel efficiency consideration is reported in Section four. The paper concludes with a summary of main findings in terms of the performance of CART models, the importance of fuel efficiency attributes in vehicle classification and suitability of the existing vehicle classification rules for energy-based applications.

Review of Existing Passenger Vehicles Classification and Associated Issues

Current passenger vehicle classification is a research result of greenhouse gas emissions (Hensher et al, 1994). Raw data were from a number of sources. A mapping process was carried out to pull all of the disparate sources of data together, to give a description of each passenger vehicle in terms of the number on register, vehicle's key physical and performance attributes, energy consumption and price by vintage.

The 1997/1998 vehicle registration database available from New South Wales Roads and Traffic Authority was used. This database contains basic physical vehicle characteristics (Make, Year, Engine Capacity, Number of Cylinders). Vehicle prices are joined to this database (Source: 1997 Glass's Guide).

Additional vehicle's attributes were collected and joined to this database. They are vehicles' fuel types and fuel efficiency (Source: Department of Primary Industry and Energy (DPIE)'s 1997/1998 Fuel Consumption Guide).

In terms of fuel efficiency, there are three measures, fuel efficiency based on city cycle (CE), fuel efficiency based on highway cycle (HE) and the combined on road fuel efficiency (RE). We use 1997/1998 DPIE's Fuel Consumption Guide for updating CE and HE attributes. The RE attribute represents a combination of both city cycle (CE) and highway cycle (HE) fuel efficiency measure. The discussion on formulating the RE attribute is detailed in Hensher (1995). Basically, the formula for calculating RE is as follows:

$$RE = 0.988462 + 0.871080 \text{ x } (0.7\text{x}CE + 0.3\text{x}HE) \quad (1)$$

Finally, we constructed a database containing 823 cases. Each case represents a unique combination of vehicle attributes including physical attributes, price, fuel type, fuel efficiency and associated class. Table 1 describes the current classification rule for assigning vehicle to specific class. There are 9 classes where micro vehicles are grouped in class 1 and four wheel drive vehicles are classified in class 9.

As indicated in Table 1, the current classification rules are based only on vehicle's physical attributes (Engine Capacity CC, Number of Cylinders CYL, Year YR, Make MAKE and Prices P). The lack of fuel efficiency attributes (such as CE, HE and RE) in the current vehicle classification raises the issue about its suitability in energy-based studies.

Figure 1 shows road fuel efficiency (RE) frequency distributions for all nine vehicle classes. With the exception of class 7 (luxury vehicles) and 9 (four wheel drive), all distributions have different means and variances. This gives an intuition that road fuel efficiency might possess an explanatory power in a vehicle classification model. If it is the case then how important fuel efficiency attributes are in comparing with other physical attributes such as engine capacity (CC), number of cylinders (CYL), year (YR), make (MAKE)and prices (P).

CART is selected as a modelling tool to represent vehicle classification rules with and without fuel efficiency variables. Basics of CART and motivation for using CART to build vehicle classification models are described in next section.

Table 1. Current vehicle classification rules

Class	Description
1	Micro (= < 4 cylinders, < 1400 cc)
2	Small (4 cylinders, 1400 - 1900 cc)
3	Medium (4 cylinders, > 1900 cc)
4	Upper Medium 1 (6 cylinders, < 3000 cc)
5	Upper Medium 2 (6 cylinders, > = 3000 cc)
6	Large (= 8 cylinders)
7	Luxury (specific makes and engine capacities). All of: Mercedes, BMW, Rolls Royce, Jaguar, Audi, Bentley, Lexus, Daimler and Eunos Plus: Honda Legend / NSX (> 3000 cc), Volvo = 2300 cc, Saab > 2100 cc
8	Light Commercial
9	Four Wheel Drive

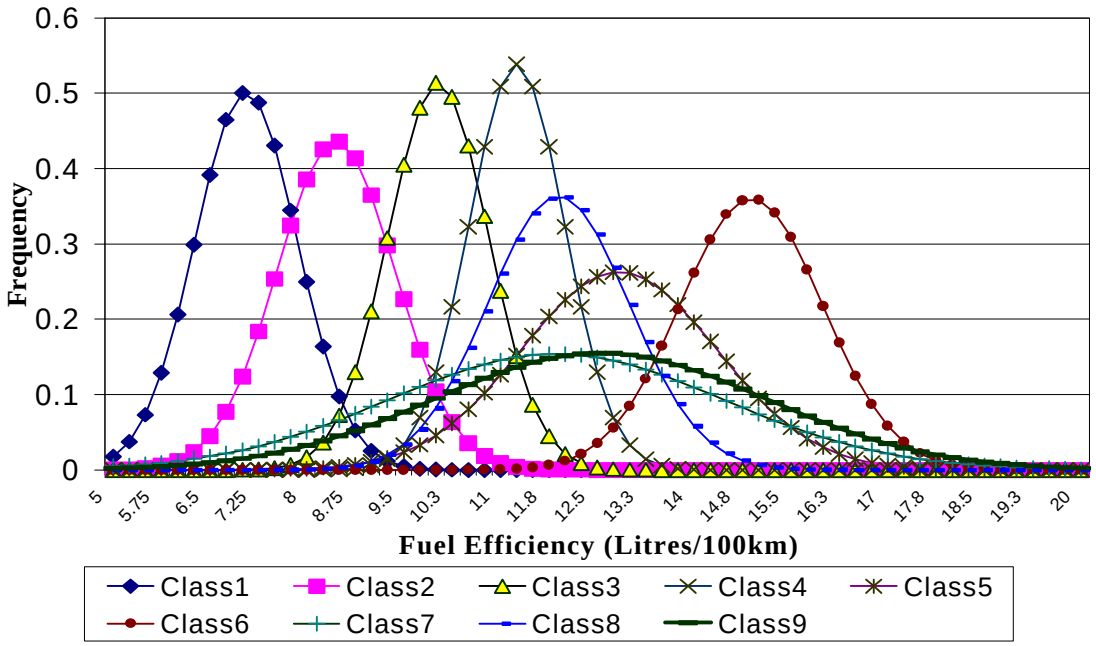


Figure 1. Comparison of road fuel efficiency frequency distributions among vehicle classes

Motivation for Using CART to Represent Vehicle Classification Problems

CART (Classification and Regression Trees) was developed by Breiman et al. (1984) and later enhanced and implemented by Steinberg and Colla (1998) to produce software package of the same name. Since then, CART has begun to interest a larger audience of researchers and practitioners among many disciplines focusing on classification problems.

For a general classification problem, one has N observations (a learning data set) of a categorical variable with levels j = 1,2, …, J, and of K independent variables, which may include both categorical and continuous variables. The objective of any classification method including CART is to use the information in the sample in some optimal way to best classify a given observation into one of the J categories or to estimate the probability that it belongs to each of these categories.

What makes CART different from the other methods is the key that CART possesses. CART uses a multi-sequential search algorithm to optimise the classification of a phenomenon and presents the results in the form of a classification (decision) tree – a significant departure from more traditional statistical procedures.

Basically, CART process is consisted of four major steps: tree growing, tree pruning, tree selection and tree testing.

Tree growing: Classification tree is generated by a set of binary recursive and iterative partitioning on data set based on the answers to questions (*splitting rules*). Questions are always presented in the form *Is CONDITION <= VALUE (eg. Is AGE <= 65)*. A

classification tree is started with a *root node.* One question forms one split at a root node and two partitions (*child nodes*) associating with YES (the cases go left) and NO (the cases go right) answers. Each child node is in turn split into other two child nodes. This process is computationally intensive but the number of splits is finite. There are at most *N* different splits for a continuous variable in a data set with sample size *N*. For a categorical variable with *L* levels, 2^{L-1} splits can be found. The key feature of this tree growing step is to find the best split at every node. CART evaluates goodness of any candidate split by using an impurity function.

There are a number of impurity functions. Gini index of diversity (Beiman et al, 1984) represents a popular impurity function. Suppose that target variable has *j=1,2,...,J* levels, the proportions of cases falling into the *J* categories are *p(i), i=1,2,...,J*, for any node *t*. The Gini measure is defined as:

$$i(t) = 1 - \sum_{i=1}^{J} p^2(i) \tag{2}$$

It is obvious that *i(t)* has its maximum value of *(J-1)/J* when *p(1)=p(2)=...=p(J)=1/J* (ie. a node which contains an equal proportion of every class is least pure) and its minimum value of 0 when one of the *p(j)=1* and all others equal 0 (ie. a node which contains members of only one class is perfectly pure).

The best split is one that maximizes the decrease in impurity.

$$\Delta(t, s) = i(t) - p_L i(t_L) - p_R i(t_R) \tag{3}$$

where *t* denotes a node and t_L and t_R are partitioned child nodes. *s* is splitting rule. p_L and p_R are the probabilities of a case going left and right. *i(t), i(t_L)* and *i(t_R)* are impurities of node *t*, t_L and t_R. *D(t,s)* denotes the improvement in impurity as the resulting of partition based on splitting rule *s*.

Once the best split is found, a search is made for the best splits of each of two child nodes. This process is then continued to grow the classification tree until no further splitting is possible (*stopping criterion*). Final tree is called the *maximal tree* and final subsets resulting from splitting are *terminal nodes* of the tree. A class character is assigned to each terminal node by the plurality rule (Breiman et al, 1984). Specifically, if

$$p(j_0 \mid t) = \max_j p(j \mid t) \tag{4}$$

then *t* is designated as class j_0 terminal node.

Tree pruning: Having grown a maximal tree by using splitting rules and stopping criteria, CART's pruning process is carried out upward to form a tree sequence, based both on minimising a linear combination of predictive accuracy and on a penalty applied to large trees (*minimal cost complexity pruning*). Misclassification cost of the tree is defined as follows:

$$R_a(T) = R(T) + a\,|\overline{T}\,| \quad (5)$$

Where α is complexity parameter. $\overline{T}$ is the number of terminal nodes indicating the tree complexity. *R(T)* is the misclassification cost of the tree.

Tree selection: Given a list of candidate trees, an optimal tree will be selected based on its minimal cost complexity measure in comparing with other trees including maximal tree.

Tree testing: The question raised in determining how accurate tree's classifiers are to test classifiers on subsequent cases whose correct classification has been observed. It should be noted that the classifier is derived from *learning sample.* Then this learning sample is used to construct classifiers and to estimate their accuracy.

In the next section, CART will be used in the development and evaluation of vehicle classification problems.

Developing and Evaluating Vehicle Classification Tree Models

Developing vehicle classification models

The development of vehicle classification trees has two aims: to produce an accurate classifier and to uncover the predictive structure of the problem. These two are not exclusive, even if the emphasis is on producing accurate classifier in constructing model.

Four classification tree models are constructed to represent current vehicle classification problem with and without fuel efficiency considerations (see Table 2). Basically, model 1 is constructed as a base model to represent the current vehicle classification scheme without fuel efficiency measures. Models 2, 3 and 4 are modified versions of base model by adding city road, highway and combined road fuel efficiency measures, respectively. All predictors are treated as continuous variables and the target variable (CLASS) is categorical, representing 9 different classes of vehicles ranging from 1 to 9. Tenfold cross-validation is used in constructing and testing vehicle classification models. The cases in learning sample are randomly divided into 10 subsets of as nearly equal size as possible. Entire learning sample is used to build maximal tree. A sequence of 10 auxiliary trees then is constructed. Each is built using all but one of 10 subsets and tested on the remaining subset.

As an illustration, Model 4 is selected to describe the CART's process of tree growing, tree pruning, tree selection and tree testing.

Table 2. Description of four vehicle classification tree models

Model	Description
1	Base model, it represents current vehicle class scheme, using CC (engine capacity, CYL (number of cylinders), P (price) and YR (year) as predictors and CLASS as target variable.
2	Modifying base model by adding CE (City road fuel efficiency variable) to predictors.
3	Modifying base model by adding HE (Highway fuel efficiency variable) to predictors.
4	Modifying base model by adding RE (Road fuel efficiency variable) to predictors.

Tree growing: The data set input to CART consisted of 823 cases of unique vehicles. This step begins with an examination of all possible splits of 823 cases into two partitions based on the values of predictors (CC, CYL, P, YR and RE) and the selection of best splits. The search for best split consists of evaluating at most 4115 splits (= 5 variables x sample size of 823 cases). The end result is a maximal tree with 66 terminal nodes. Tree pruning procedure was then applied to prune this maximal tree.

Tree pruning: A sequence of 37 trees is formed as a result of pruning maximal tree. Table 3 lists 13 of them. On this table, tree 1 is the maximal tree, tree 13 represents minimal cost tree and tree 20 is an optimal tree. The relative cost for testing (cross validation or resubtitution) is the sum of misclassification rates of all terminal nodes for a tree classifier based on the method of testing. The resubstitution cost *R(T)* (in Table 3) of a tree represents the sum of misclassification rates of all terminal nodes if learning data is reused for testing the model. The cross validation cost $R^{CV}(T)$ is estimated as the average performance of 10 test samples based on 10 fold cross validation. This method represents an efficient use of available data since each case is used in constructing the tree and each case is exactly used once in a test sample.

The complexity parameter α is defined in Equation 5 above. The higher number of nodes will result in lower value for complexity parameter α. With 66 nodes, maximal tree (tree 1) has α equal to 0. As number of nodes decreases, complexity parameter increases.

Tree selection: The cross-validation cost initially decreases rapidly followed by a long, flat range, then increases gradually as tree grows. The minimum is unstable and occurs somewhere in the flat range. Therefore, we introduce 1 SE (Standard Error) to overcome the instability. The optimal tree is tree 20 (Table 3) corresponding the estimated minimal cross-validation misclassification cost (in tree 13) plus 1 SE. The optimal tree has 27 terminal nodes with cross-validated relative cost 0.134 and resubstitution relative cost 0.105.

Figure 2 shows the optimal tree for representing Model 4 of vehicle classification problem. On this figure, there are two types of nodes: splitting nodes (represented by hexagons) and terminal nodes (represented by quadrilaterals). Splitting nodes are nodes that can be further split to two child nodes. The root node is also a splitting node.

Table 3. Tree sequence for vehicle classification model

Terminal		Cross-Validated	Resubstitution	Complexity
Tree	Nodes	Relative Cost $R^{CV}(T)$	Relative Cost $R(T)$	Parameter
1	66	0.135 ± 0.014	0.059	0.000
13*	38	0.126 ± 0.014	0.083	0.001
20**	27	0.134 ± 0.014	0.105	0.002
28	11	0.195 ± 0.015	0.180	0.006
29	10	0.206 ± 0.015	0.196	0.014
30	9	0.232 ± 0.014	0.216	0.018
31	8	0.261 ± 0.014	0.246	0.027
32	7	0.327 ± 0.009	0.300	0.048
33	6	0.382 ± 0.003	0.380	0.071
34	5	0.468 ± 0.005	0.503	0.109
35	4	0.642 ± 0.008	0.627	0.110
36	2	0.813 ± 0.010	0.875	0.110
37	1	1.000 ± 0.000	1.000	0.111

Note: (*) minimum cost tree (**) optimal tree

Terminal nodes are nodes that cannot be further split into child nodes. At splitting nodes, splitting criterion was given as well as node index, assigned class and number of cases.

Starting from the root node at the top of the tree in Figure 2, we can see that the splitting rule based on number of cylinders (CYL) where 462 cases go left (node 2) when the number of cylinders is less than 5. It should be noted that CART uses mid-point splitting between CYL = 4 and CYL = 5. Therefore the splitting rule is displayed as CYL <=4.5. Those 462 cases at node 2 are further split based on road fuel efficiency (RE <= 8.127). There are 183 cases satisfying RE <=8.127. These cases are sent to the left node (node 3). Based on vehicle prices, these 183 cases are then split into two child nodes: terminal node –1 and node 4. At terminal node –1, there are 47 cases classified as class 1. The tree keeps growing from node 4 to other nodes. Other part of the tree follows similar mechanism: binary partitioning until no further split is found. In total, there are 26 split nodes forming 27 terminal nodes.

Tree testing: The performance of classification tree is best reviewed in prediction success. It is possible that a classification tree predicts the learning sample well, while not necessarily performing well on new data. It is not easy to draw new independent data from same distribution as learning sample and use it for testing purpose. We use ten-fold cross-validation testing sample to assess the predictive power of generated tree. The cross-validation testing represents an estimate of the results that would occur if the tree was applied to new data drawn from the same distribution as the learning data.

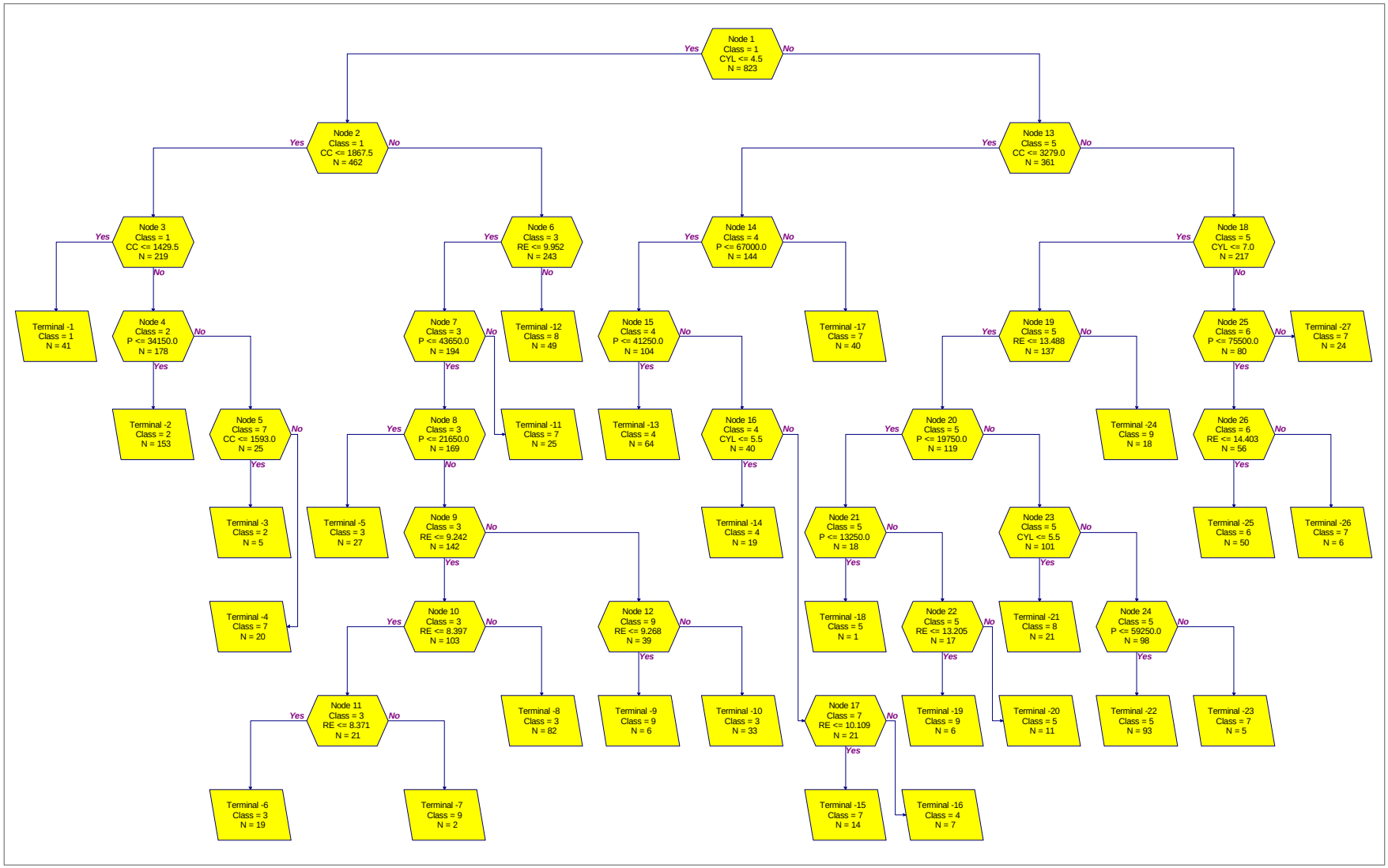


Figure 2. A typical vehicle classification tree

The notation of prediction success was first introduced by McFadden (1979) and is developed in Hensher and Johnson (1981). Table 4 gives the prediction success of 10-fold cross validation testing. Cases appearing on the diagonals of the matrix correspond to correct classification, while off-diagonal entries represent misclassification. The sum of off-diagonal entries is overall misclassifications. While the contrast of predicted total to actual total for each class gives a rough idea how well does the tree perform, 'Correct' (Table 4) precisely indicates proportion successfully predicted for each class. We also give the Success Index (Hensher and Johnson 1981:54) for each case, which denotes the fraction by which the percent correct exceeds what would be expected on the basis of chance alone. As a whole, classification tree for cross-validation testing samples has unweighted overall correct of 0.885 and misclassified cases of 95, compared to overall correct of 0.917 and misclassified cases of 68 for learning sample.

Evaluating vehicle classification tree models

In this section, we compare the existing vehicle classification scheme (Model 1) with other the three classification models (Models 2, 3, and 4) that include fuel efficiency measures. The model comparison task has two aims. First, we investigate the suitability of fuel efficiency measures as criteria in vehicle classification problems represented by Models 2, 3 and 4. Second, we identify the relative importance of fuel efficiency attributes in comparing with other physical attributes such as engine capacity CC, number of cylinders CYL, year YR, make MAKE and prices P.

Table 4. Prediction success of vehicle classification tree

Actual Class	Predicted Class									Actual
	1	2	3	4	5	6	7	8	9	Total
1	35	0	0	0	0	0	0	0	0	35
2	3	154	0	0	0	0	9	0	0	166
3	2	0	145	0	0	0	4	4	3	158
4	0	0	0	76	0	0	2	3	1	82
5	0	0	0	1	94	0	0	1	2	98
6	0	0	0	0	0	48	0	0	0	48
7	1	6	5	6	5	1	110	10	0	144
8	0	0	0	2	4	2	2	38	5	53
9	0	0	4	2	2	0	0	3	28	39
Predicted Total	41	160	154	87	105	51	127	59	39	823
Correct	1.000	0.928	0.918	0.927	0.959	1.000	0.764	0.717	0.718	
Success Index	0.957	0.726	0.726	0.827	0.840	0.942	0.589	0.653	0.671	

Cross-Validation: Number of Misclassification:95, Total Correct: 0.885
Learning Sample: Number of Misclassification: 68, Total Correct: 0.917

Investigate the suitability of fuel efficiency measures in vehicle classification problems: Table 5 summarises the basic particulars in comparing the performance of these four models. In term of tree structure, the three fuel efficiency based models (Models 2, 3 and 4) yield less complex structure than the base model (Model 1). Model 2 (incorporating city cycle fuel efficiency CE) and Model 4 (incorporating road fuel efficiency) perform better than the base model (without any fuel efficiency measure) on the ground of training, testing error, misclassification and total correct. Model 3 which include highway cycle fuel efficiency measure HE comes quite close to the base model on every particulars. The implication of this finding is that fuel efficiency measures might be considered as predictors in vehicle classification problems.

The suitability of fuel efficiency measure is further demonstrated by contrasting predictive performance of Model 1 (base case) and Model 4 (base plus RE attribute) (see Figure 3). In Figure 3 the predicted classes by Model 1 and Model 4 are plotted next to actual classes from the sample. Model 4 outperforms Model 1 nearly for every vehicle class except cases in class 1 and class 7.

Class 1: Micro (< = 4 cylinders, < 1400 cc)

Both Models 1 and 4 predict slightly more members of this class. The inclusion of road fuel efficiency in Model 4 allows some class 2 vehicles with superior fuel efficiency moving into class 1.

Class 2: Small (4 cylinders, 1400 - 1900 cc)

Model 1 significantly overpredicts number of vehicles in this class. Misclassification of 5 cylinder vehicles is one possibility, but minor. Misclassification of small 4WDs (four wheel drive) would explain some variations, especially where 4WD is not a major factor in the vehicles design eg Honda CRV and Volvo AWD. Number of vehicles assigned to this class by Model 4 is slightly less than actual. One possible reason might be the fact that the inclusion of road fuel efficiency in Model 4 enables frontal area, body type and streamlining effects to be incorporated. The extent of variation from Model 1 suggests a wide range in vehicle outcomes for common attribute specifications - possibly extra costs for higher quality produce significant differences in on-road performance. Treatment of rotary engined vehicles with swept volumes of 1200 to 1310 cc but nominal capacities of 1680-1834 cc (1.4) is another minor source of variation, which Model 4 would better predict. It seems that recent market innovation has been concentrated here – eg. new Korean entrants.

Class 3: Medium (4 cylinders, > 1900 cc)

Both Model 1 and Model 4 are quite close to actual classification. It may reflect large size and technical maturity of this segment, where the market has evolved to stable relationships between vehicle volume and engine size, gearing, variations of different body shapes etc. This stability reflected in minor differences in fuel consumption between specification based classification, and fuel consumption based classification.

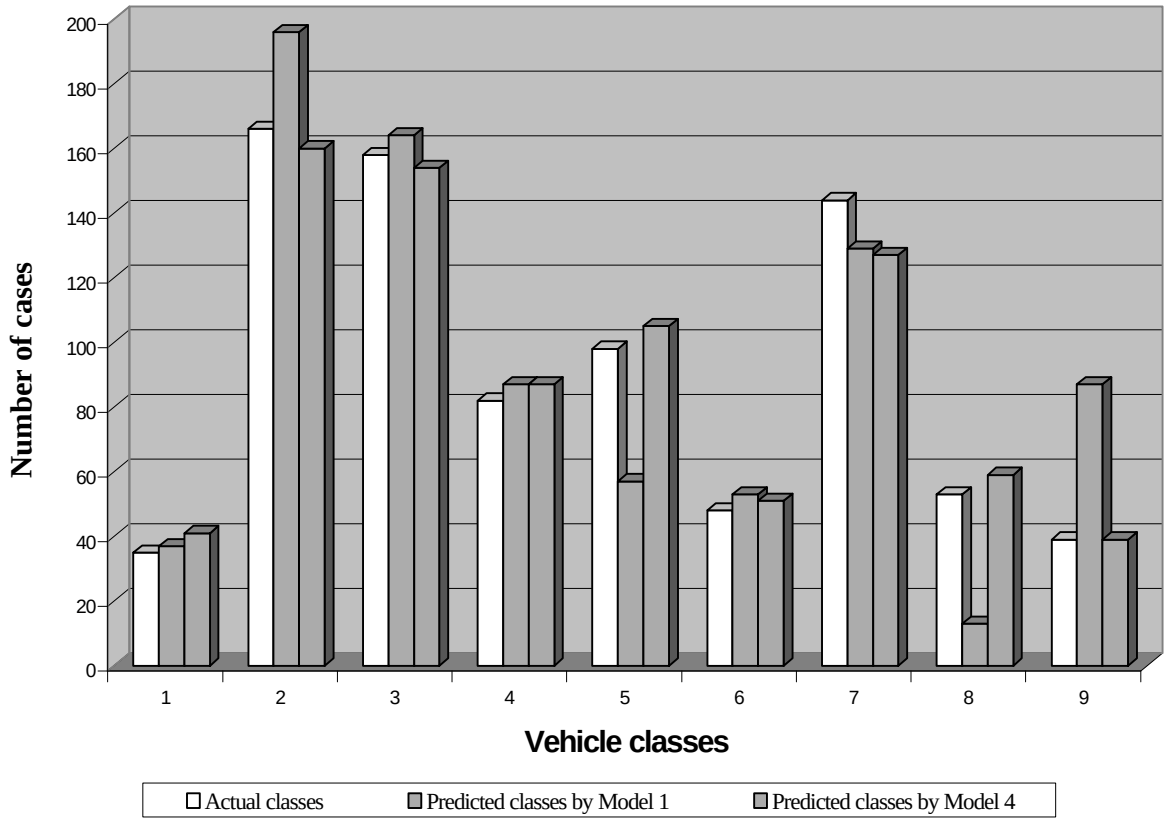
Table 5. Comparison of model performance

Particulars	Model 1	Model 2	Model 3	Model 4
Tree structure	30 nodes	18 nodes	21 nodes	27 nodes
Training cost/error	0.240	0.192	0.264	0.134
Testing cost/error	0.283	0.202	0.275	0.105
Number of misclassified cases	173	133	174	95
Total correct	0.790	0.838	0.789	0.885

Class 4: Upper medium 1 (6 cylinders, < 3000 cc)

Both Model 1 and Model 4 produce very close classification levels, slightly above actual classification. Many Class 4 vehicles are developed from Class 3 types to which a six cylinder version has evolved, eg. Mitsubishi Magna/Verada. Small six cylinder engines are generally just under 3 litres in size, so there is a low level of variation in vehicle or engine sizes (eg. Volvo and Volkswagen 5 cylinder engines are approx 2.4 litres, equivalent to a large 4 cylinder).

Figure 3. Predictive performance of model 1 and 4 against actual vehicle classes



Class 5: Upper medium 2 (6 cylinders, > = 3000 cc)

Distinct differences in Model 1 and Model 4 performance here. Model 1 underpredicts by about a third, while Model 4 slightly overpredicts. Possible reasons for this are greater variation in engine size, from just over 3 litres to 4.1 litres engine capacity. In addition, standard family sizes have stopped growing with 3.8 litres for Holden and 4.0 litre for Ford current standards. Earlier 3.3 and 4.1 litre engines remain in the vehicle fleet.

Class 6: Large (= 8 cylinders)

Both Model 1 and Model 4 produce good results, although Model 1 overstates to a greater degree. 4.2 to 5.8 litre capacity engines are extant in the vehicle fleet, but this market segment has been declining so existing engines continue to be used. The better prediction by Model 4 probably reflects that development effort has generally been directed towards emissions compliance.

Class 7: Luxury (specific makes and engine capacities).

Model 1 underpredicts this class by about one sixth, with Model 4 fractionally provides smaller estimate. These vehicles allocated to the luxury classes display a much greater variation in engine capacity ostensible purpose and engine technologies. Because vehicles are uniquely classified by make into this class lower model performance is likely allocated in this class. The slightly lower performance of the fuel -efficiency model probably reflects the great dispersion between 1800 cc to 6750 cc engines and other vehicles with 74 kW to 410 kW power ratings. While vehicle price (P) puts many sport vehicles in this class they would expect to have high power to weight ratios than highly specified luxury orientated vehicles. There are also significant variations in technology levels, with 4 to 12 cylinder engines and other advanced engineering applications.

Class 8: Light commercial

Model 1 predicts only about a fifth of vehicles in this class, while Model 4 predicts slightly more. Because this class involves light commercial vehicles power-to weight ratios are likely to be lower. Relativities will also vary between passenger cars derived vehicles and separately developed vans etc. Significant variation in this class, with one model likely to have height and length options, while slightly heavier but more fuel -efficient diesel engines are widely available and used. This class is also likely to have greatest variation between fully laden and unladen fuel consumption. Model 4' s on-road fuel consumption element offers integration of most such differences.

Class 9: Four wheel drive

Model 1 predicts about double the number in this class. Because the additional weight and volume of the transfer case, differential(s) drive shafts, large wheels etc higher fuel consumption is to be expected for these 4WDs. There is also the possibility of a bi-modal distribution, with full size models like Land Cruiser, Patrol and Jeep having distinct performance from lighter, lifestyle orientated 4WDs like Rav, Sierra, Freelander etc. Model 4' s on-road fuel consumption element offers integration of most such differences. Model 1 is likely to be sensitive to whether a vehicle is considered a proper 4WD eg, the Mercedes Benz M class or Volkswagen Synchro Kombi variants.

Identify the relative importance of fuel efficiency attributes in comparing with other vehicle attributes: Each variable in vehicle classification tree has an importance score

based on how often and with what significance it served as primary or surrogate splitter throughout the tree.

The notations of primary and surrogate split were developed in Breiman et al (1984). At every node, the best split is searched by first finding the best split for a specific variable. Then this search is repeated over all other variables. The primary split is the best split of a node of all measurement variables measured in reducing impurity in two child nodes. A surrogate is a split that splits in a fashion similar to the primary. It is a variable with possibly equivalent information which is useful in revealing the structure of information in variables. The capability that a surrogate split mimics the primary split is expressed as predictive association between primary and surrogate split which is the reduction in mismatch between primary and surrogate splits relative to primary mismatch. If match is perfect then surrogate mismatch is 0 and association is 1. Association could be negative and will be for most variables. A split qualifies as a surrogate only if association is greater than 0. Surrogates are ranked in order of association. If a primary split is missing, the first surrogate is used in splitting. If the first is also missing, then second is used and so on.

Several alternative methods of computing importance are available in CART software (Steinberg and Colla, 1998), including ignoring the contributions of surrogate splitters, discounting them by their association or a geometric factor, and only considering the top N surrogates for each node (rather than all available surrogates for each node). The scores reflect the contribution each variable makes in classifying or predicting the target variable, with the contribution stemming from both the variable's role in primary splits and its role as a surrogate splitter.

With these notations in mind, we then investigate the functions of on road fuel efficiency (CE) in the tree structure. As shown in Table 6, it is used as the primary split in 9 of 26 splitting nodes of the optimal tree. In all these nodes, engine capacity is always qualified as the first surrogate split, except in the node 19 and 22, where there is no available surrogate. This finding indicates that on road fuel efficiency (CE) is the best criterion in vehicle classification in these nodes. Engine capacity could be important but its importance is masked by fuel efficiency in some of these nodes. This is indicated in splitting improvement. If surrogate is used in splitting instead, improvement is very poor in the node 6, 9, 11 and 12.

Road fuel efficiency is qualified as a surrogate split in 13 of 26 splitting nodes (Table 7). The importance of fuel efficiency could be masked by primary split in these nodes. The relative high improvement can be reached in nodes of 1, 2, 5, 13, 15, and 25 if fuel efficiency is used as splitting variable.

Table 6. Road fuel efficiency (CE) as a primary split

Node	Split Value	Improvement	First Surrogate Split		
			Variable: Value	Association	Improvement
6	9.952	0.151	CC:2413.5	0.387	0.101
9	9.242	0.027	CC:2423.0	0.327	1.12E-05
10	8.397	0.004	CC:2506.0	0.113	3.04E-07
11	8.371	0.016	CC:2174.5	0.753	0.000
12	9.268	0.035	CC:2404.0	0.547	2.85E-05
17	10.109	0.018	CC:2857.5	0.703	0.014
19	13.488	0.108	*		
22	13.205	0.028	*		
26	14.403	0.018	CC:5751.5	0.547	0.017

* indicates there is no available surrogate

Table 7. Road fuel efficiency as a surrogate split

Node	Primary Split		Road Fuel Efficiency as a Surrogate Split		
	Variable:Value	Improvement	Split Value	Association	Improvement
1	CYL:4.5	0.602	9.346	0.675	0.502
2	CC:1867.5	0.371	8.127*	0.746	0.346
3	CC:1429.5	0.229	6.921*	0.626	0.097
5	CC:1539	0.006	7.487*	0.800	0.004
8	P:21650	0.052	7.278	0.030	2.31E-04
13	CC:3279	0.289	10.470*	0.614	0.194
14	P:67000	0.064	7.996	0.026	0.002
15	P:41250	0.014	11.703	0.045	0.008
16	CYL:5.5	0.026	10.109	0.490	0.005
18	CYL:7.0	0.295	11.324	0.060	0.106
20	P:19750	0.068	12.683*	0.282	0.013
21	P:13250	0.029	10.252	1.000	0.001
25	P:75500	0.063	14.011	0.200	0.032

* indicates that road fuel efficiency is the first surrogate split

Table 8 summarises the variable importance among the four models. Engine capacity (CC) is the most important classifier in the base model (Model 1), Model 2 and Model 3. Among the three fuel efficiency measures (city cycle CE, highway cycle HE and road RE), the road efficiency RE ranked first in Model 4 even higher than engine capacity (CC). The city cycle fuel efficiency (CE) used in Model 2 ranked second after engine capacity (CC). The highway cycle fuel efficiency (HE) used in Model 3 ranked fourth after engine capacity (CC), price (P) and number of cylinders (CYL).

We reveal the relationship between the vehicle fuel efficiency and engine capacity by

developing a new classification model tree by excluding engine capacity (CC) as a predictor just to test its impact to the model performance.

It is possible that road fuel efficiency (CE) and engine capacity (CC) are highly correlated. Consequently, the relative importance of one variable is masked or covered by another variable. To explore this possibility, we construct another tree by excluding the engine capacity as a predictor. The optimal tree has 25 terminal nodes. To make two trees comparable, complexity parameter of second tree is set to equal of previous tree, i.e. 0.002. The accuracy of second tree decreases significantly (Table 9). A number of 210 cases are misclassified in 10-fold cross-validation testing with relative cost of 0.256. The resubstitution relative cost is 0.220, more than twice of the original tree of Model 4 (including CC). The total correct rate is correspondingly reduced to 0.745.

Table 8. Comparison of the variable importance among the four models

Ranking	Model 1	Model 2	Model 3	Model 4
1	CC * (100)	CC (100)	CC (100)	RE (100)
2	P (89.39)	CE (94.61)	P (83.20)	CC (94.71)
3	CYL (67.31)	CYL (78.06)	CYL (71.10)	CYL (74.63)
4	YR (8.78)	P (75.00)	HE (66.89)	P (73.63)
5		YR (12.20)	YR (8.63)	YR (9.51)

Note: (*) Name of variable used in model and relative important score in brackets.

Table 9. Tree comparison of two versions of Model 4 (with and without engine capacity CC)

Item	Include CC as a Predictor	Exclude CC as a Predictor
Nodes of Best Tree	27	25
Cross-Validation Relative Cost	0.134	0.256
Resubstitution Relative Cost	0.105	0.220
Number of Misclassification	95	210
Total Correct	0.885	0.745

Conclusions

Two features of CART have proven to be useful in classifying vehicles taking into account physical and or fuel efficiency attributes. First CART could automatically analyse data and CART could separate relevant from irrelevant predictors. Second CART could yield relatively simple and easy to comprehend models. These features are desirable for multi-dimensional classification problem such as vehicle classification problem. Normally, the way a number of variables in classification problem may not be well understood by data analyst. By examining all possible variables for the best classifiers and optimal tree, complex interactions among the variables are easily identified.

Engine capacity represents the most important attribute without the presence of fuel efficiency variable. On road fuel efficiency measure was ranked higher than engine capacity in its importance in structuring vehicle classification model. However, engine capacity should still be considered relevant predictor in vehicle classification model. In other words, the performance of vehicle classification model would deteriorate if engine capacity variable is not included in classification model's structure.

We encountered the problem of accuracy of cross-validation due to the sample size of testing data. Numbers of cases in class 1, 6, 8 and 9 are 35, 48, 53 and 39 respectively. If these cases are randomly divided into sub-sample of about 10 percent and randomly combined, these under-represented cases might be highly unevenly distributed among sub-samples. To overcome this problem, stratification on categorical target variable CLASS was used. Further research will be carried out in using different methods for specifying splitting rules and computing variable's importance. Currently, the condition in splitting rule (*Is CONDITION <= VALUE*) is a measured variable. More complex conditions can be used such as Boolean or linear combinations.

Acknowledgment

The research and development reported in this paper is supported under the Australian Research Council Research Centres program. Valuable comments and supports from Prof. David Hensher, Messrs Cam Ngo and Kirk Bendall at Institute of Transport Studies, the University of Sydney and the referee are acknowledged.

References

Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J. (1984) *Classification and Regression Trees* California: Wadsworth International Group

Department of Primary Industries and Energy (1997) *Fuel Consumption Guide For Buyer of New Cars, Four-Wheel Drives and Light Commercials*, Canberra: Australian Government Publishing Service.

Glass's Dealers Guide Pty Ltd (1997) *Glass' s Dealers Guide to Passenger and Light Commercial Vehicle Values*, Melbourne: Glass's Dealers Guide Pty Ltd.

Hensher, D.A. & Johnson, L.W. (1981) *Applied Discrete Choice Modelling* London: Croom Helm

Hensher, D.A, Battellino, H., Milthorpe, F. & Rainmond, T. (1994) *Greenhouse Gas Emissions and the Demand for Urban Passenger Transport: Data Requirement, Documentation and Preparation*, Institute of Transport Studies, The University of Sydney (Unpublished Report)

McFadden, D. (1979) Quantitative methods for analysing travel behaviour of individuals, in Hensher, D.A. & Stopher, P.R. (eds) *Behavioural Travel Modelling*, Croom Helm, London.

Steinberg, D. and Colla, P. (1998) CART – Classification and regression tress, San Diego, CA: Salford Systems.

INSTITUTE OF TRANSPORT STUDIES

The Australian Key Centre
in Transport Management

The University of Sydney
and Monash University