American Journal of Applied Sciences

Original Research Paper

A Comparison between Different Data Mining Algorithms in Freight Mode Choice

Amir Samimi, Hesamoddin Razi-Ardakani and Amir Nohekhan

Department of Civil Engineering, Sharif University of Technology, Azadi Avenue, Tehran 11365-8639, Iran

Article history
Received: 01-11-2016
Revised: 19-12-2016
Accepted: 20-12-2016

Corresponding Author:
Amir Samimi
Department of Civil Engineering, Sharif University of Technology, Azadi Avenue, Tehran 11365-8639, Iran
Email: amir.samimi@gmail.com

Abstract: This research aims to study application of support vector machine algorithm, artificial neural networks and five different types of decision trees in predicting mode choice of freight transportation. Performance of these models has been compared with log it model which is one the most prevalent statistical models in the field. Effect of factors such as cargo weight, distance, type and characteristics of commodity has been studied in process of modelling mode choice which is rail and road. In this regard, data gathered in the United States, is used and similarities and advantages of the models are described in details. Results indicated that cost-sensitive support vector machine is the best method in predicting shipment mode choice. After this method, stand C5 decision tree and artificial neural network. The most important variables in determining shipment mode choice of firms are respectively weight, great-circle distance between origin and destination, commodity type, compound impedance factor of rail and truck and containerized condition of the shipment to be moved.

Keywords: Freight Mode Choice, Logit Model, Decision Trees, Support Vector Machine, Artificial Neural Network

Introduction

Transporting goods in road network causes multiple drawbacks. The most important disadvantages to be mentioned are safety reduction, increase in travel time, increased emissions and energy consumption and destruction of infrastructures (Wisetjindawat *et al*., 2015). On the other side, road transportation network has a pivotal role in economic growth of a country and also propelling industries related to this type of transportation. Facilitating the process of freight transporting has a direct impact on a country's economy; the competition between rail and road plays a major role in assignment of funds and resources to related segments (de Dios Ortúzar and Willumsen, 2001). Therefore behavioural analysis of firms and identification of factors affecting their decisions can greatly benefit efficient policy in this scope.

The main challenges in discussing issues of freight transportation are the volume of goods moved, mode choice, route choice and scheduling goods movement. Mode choice can be regarded as one of the most substantial decisions of each firm involved in a supply chain (Samimi *et al*., 2011). Generally stating, in studying behaviour of firms in field of freight mode choice there are much more complicity than mode choice of commuters. While in commuters mode choice it is dealt with people each having their own characteristics, in freight transportation it is dealt with firms each of them in one or more supply chains and their respected decisions are usually made joint, resulting in complex behaviour. However, procedures which are being used in commuter's mode choice can be applied to freight transportation with modifications. Employing models able to examine non-linear relations between independent and dependent variables may be regarded as a suitable option in studying freight mode choice decisions.

According to past studies, decision trees and support vector machine have not been employed yet in order to predict freight mode choice decisions. Few studies explored performance of data mining algorithms in this context. The first research in this area was conducted by Abdelwahab and Sayed (1999). They compared performance of logit and probit models with artificial neural network in predicting freight mode choice (truck and rail). In their study, models were compared in four conditions; three conditions, regarding portion of data used for model estimation, includes model estimation using the whole variables in data, significant variables in statistics model and variables with the most weight in input nodes of the artificial neural network. Criterion of comparison was aggregate precision of classification.

Results indicated relatively equal performance of models. Also based on output probability of probit model and output number resulting from artificial network (which is in interval between 0 and 1), by changing the threshold of 0.5 for classification and increasing confidence interval, outcomes of these two models were compared. In this case artificial network had shown much better performance. Sayed and Razavi (2000) compared performance of neural-fuzzy algorithm with neural networks and logit model on the same data. Performances of three models from a classification accuracy view were estimated to be equal. As for fuzzy model's estimation, there is no obligation to choose significant variables or satisfying statistics assumptions, applying this model on mode choice is stated to be straightforward. The most recent study in this field has been done by Tortum *et al*. (2009). They generated models of intercity freight mode choice in four different countries. In this study performance of logit model, neural and neural-fuzzy networks are compared. Different criteria for exploring performance of those models is presented and based on them, the best models are introduced to be neural-fuzzy networks and neural networks. In a study over Chukyo Metropolitan Area, Japan by Wisetjindawat *et al*. (2015) using a logistic regression model for imbalanced data it was found that establishing a new Off-Rail Station can increase rail share from 8.5 to 9.2% and a subsidization can increase rail share from 15.8 to 17.9% for expected impacted industries. Combes and Tavasszy (2016) using data on shipments sent from or received in France based on the economic order quantity and the concept of total logistic costs found that the commodity flow rate contributes significantly to model's explanatory power.

In other contexts of freight demand analysis, application of data mining techniques can be observed. Celik (2004) compared neural network models application instead of gravity models in demand distribution. His study revealed that based on ordinary least squares method, neural network performs better. Nijkamp *et al*. (2004) employed neural network and logit model in order to estimate interregional freight flow in Europe. Analysis of goods flow rate based on varying toll illustrated more sensibility of logit model to input variables and results in more rational solutions.

In studies in the field of commuter's mode choice, application of eclectic algorithms is observable. Xie *et al*. (2003) estimated commuter's mode choice of work trips by decision trees, neural networks and logit model. Independent variables were five different modes. Results indicated that in prediction, neural network has the best precision; also decision tree outperforms logit model. Rashidi and Mohammadian (2011) predicted trip generation and mode choice of households using Chi-squared Automatic Interaction Detection (CHAID) decision trees hierarchically. Testing results with real data showed good performance of the algorithm. In recent years, generally, in various fields of transportation, use of decision trees to classify data is observable (Chang and Wang, 2006). Zhang and Xie (2008) studied prediction robustness of support vector machine algorithm with artificial neural networks and multinomial logit. Results indicated that in training stage, neural network had the best fitness; while in test data, support vector machine and logit model produced higher precision. Moons *et al*. (2007) compared support vector machine and Classification and Regression Tree (CART) analysis with logit model in predicting mode choice of commuters. Data was comprised of 1025 observation and based on travel type had been divided to three categories. Results of applying those models for each category indicates the best prediction precision of support vector machine in two categories. Yet, logit model performed better in prediction of the smaller class in a category with uneven distribution of objective variable.

According to the review of different studies, there are still many problems in the way of using econometric models for behavioural evaluation of freight mode choice. Also recent studies, specifically in prediction context, indicated that data mining algorithms are more accurate than others (Lu and Kawamura, 2010). Limited number of studies in this area and the lack of a comprehensive evaluation of performance of data mining models in the current issue require further studies. This study aims to evaluate and compare the ability of various algorithms in goods mode choice.

Materials and Methods

Data used in this study is based upon an online survey that was conducted in April and May 2009 by Samimi *et al*. (2010). Characteristics such as origin, destination, transportation mode, type, value, weight and volume of the commodity, cost and entire time of commodity movement were obtained for 881 freight movements. Also, some other information including number of employees, industry type, location, warehousing situation and potential use of each freight transportation mode, were collected. Significant variables of the model provided by Samimi *et al*. (2012) for predicting mode choice of commodity along with other variables which can be used for prediction, is presented in Table 1.

C5 Decision Tree

Decision trees are non-linear and non-parametric methods of data segregation (Kass, 1980). In these methods, data is recursively divided in a manner that each subset covers a homogeneous condition of objective variable.

Table 1. Explanatory Variables obtained from Samimi *et al*. (2012)

Variable	Definition
Mode	Rail or any combination of rail with other modes=1, Otherwise=0
GCD	Great-circle distance (1000 miles)
Weight	Shipment weight (1000 pounds)
Highway impedance	H
Rail impedance	R
Impedance	=EXP(H/R)
Containerized	Containerized shipment=1, otherwise=0
Commodity	Agricultural, chemical, pharmaceutical, gravel, natural sands, cement, machinery, metal, mixed freight, or prepared foodstuffs=1, otherwise=0
Value	Shipment value (USD)
Perishable	Perishable commodity=1, otherwise=0
Consolidation center	A consolidation center is used=1, otherwise=0
Distribution center	A distribution center is used=1, otherwise=0
Warehouse	A warehouse is used=1, otherwise=0
Decision-maker	A 3PL has made shipping decision=1, otherwise =0
Shipment type	Agricultural=1, chemical or pharmaceutical=2, minerals=3, electronic=4, gravel sand or cement=5, machinery=6, compound freight=7, motor vehicles and parts=8, food requirements=9, wood and paper=10, others=11

In each division of the tree, effect of all input variables on the objective variable is evaluated (Breiman *et al*., 1984). The decision tree is formed when the incursive process is completed. In forming a tree, three factors are to be determined: (1) criteria of dividing each node to son nodes, (2) measure of classification accuracy and (3) criterion for choosing final tree for classification. Based on these, various methods are provided to form a tree (Loh and Shin, 1997).

The C5 algorithm is a modification of C4.5 and ID3 trees (Quinlan, 1993). Dividing each node is computed based on information gained. This measure is used to choose the frail variable in the process of tree formation (Kotsiantis, 2007). Homogeneousness of samples in each node is defined by entropy measure. In order to compute information gain, entropy needs to be calculated first. If the objective variable has different *c* values, entropy of *S* is dependent on *c* class, which is obtained from Equation 1 (Kotsiantis, 2007):

$$Entropy(S) = -\sum_{i=1}^{c} p_i \log_2^{p_i} \tag{1}$$

where, p_i is a proportion of *S* belonging to class *i*. Information gained indicates expected decrease in entropy. Entropy demonstrates pureness of data in a choice and information gain specifies effect of a variable on classification. Information gain of (*S,A*) related to variable *A* dependent on *S* data is calculated as bellow (Kotsiantis, 2007):

$$Gain(S,A) = Entropy(S) - \sum_{v \in Value(A)} \frac{|s_v|}{|s|} Entropy(s_v) \tag{2}$$

where, *value(A)* is the whole feasible values of variable *A*, S_v is a subset of *S*, equals to *v* for variable *A*. The first term of the above equation is entropy of *S* in initial condition and the second term is expected entropy after dividing based on variable *A*. In each grown branch of the tree each variable is appeared only once. Growing of a tree continues to a point that the whole variables are appeared in a branch or the whole data in a node are dependent over one category. As information gain for variables with great values is biased, Kotsiantis (2007) presented proportion of gain to prevent such an error. It can be computed for variable *A* using equation bellow (Kotsiantis, 2007):

$$Gain\ Ratio = \frac{Gain(A)}{Entropy(A)} \tag{3}$$

Artificial Neural Network

Artificial Neural Network (ANN) is a mathematical model, which can establish and model non-linear relationships between explanatory and dependent variables (Hensher and Ton, 2000). This model, with multilayer perceptron structure, generally, is formed of three layers, each layer composed of processing units called neurons (cells, units or nodes).The first layer is the input layer, comprising of input data vector. In this layer, no process would be performed. The last layer of each network is the output layer, containing the output mapped vectors (Abdelwahab and Sayed, 1999). Also, every perceptron consists of some intermediate layer named hidden layer. Normally, every layer's neurons connect directionally to all of the neurons in adjacent layers, with a certain vector, transferring data between neurons. These connections have specific weights multiplied by the transferring data between neurons. After being weighted and transformed by a function (determined by the network's designer), the activations

of these neurons are then passed on to other neurons (Tortum *et al*., 2009). Each neuron, receives weighted outputs $W_{ij}X_i$ from the previous layer neurons and their aggregate, Net_j is the input for the neuron. Net_j can be calculated as follows (Pal and Mitra, 1992):

$$Net_j = \sum_i W_{ij}X_i + b_i \quad (4)$$

where, W_{ij} is the connection weight between node i and j. X_i is the outcome from node i and b_i is the bias of node i.

The neurons pass the received input through an activation function (threshold) to produce output. Activation functions have various types such as binary function, sigmoid, hyperbolic tangent, linear and Gaussian. The most common function in this context is the sigmoid function, as follows (Pal and Mitra, 1992):

$$Y_i = f\left(Net_j\right) = \frac{1}{1+\exp^{-(net_j)}} \quad (5)$$

The training of the ANN means calculating the various connection weights. In order to train the ANN learning algorithms each containing an input vector and a corresponding output vector is used (Celik, 2004). The number of neurons in the input and output layers, respectively are equal to the number of input and output vectors. The lack of a specific relationship for calculating the number of hidden layers and their neurons leads to testing different structures. Neural networks generally do not have good extrapolation performance which should be considered in choosing the learning algorithms. Therefore, before using neural network, the algorithms are divided in two categories namely training algorithm and testing algorithm where the training algorithm should cover the entire data as much as possible (Tortum *et al*., 2009). Obviously more training can increase generalizability of the network. Although training is a process that requires a long time, but after generalization, provides quickly the corresponding output for each input. Generally, the learning of neural network categorizes in two paradigms namely, supervised learning which assigns a certain output to any input and unsupervised learning, in which the output is unspecified. In this study artificial neural network with multilayer perceptron has been used (Pal and Mitra, 1992).

Support Vector Machine

The main idea of this method is presented by Vapnik (1995). Explaining this method is performed based on Izenman study (2008). Support Vector Machine (SVM) is designed for large scale systems with sparse data (data with few training samples). This method is known to be a model with no parameters. The approach in this method keeps training error in the range of zero or an acceptable level along with minimizing estimation error. In short, SVM utilizes a hyper plane to establish a model with maximum bounds. For this purpose, in training stage of this algorithm, dividing data is transformed to a constrained non-linear optimization problem which is intrinsically a second order programming problem. In situations where data is not dividable, SVM with the aid of Kernel functions takes data into an input space with higher dimensions and then attempts to resolve new data linearly. For the purpose of finding a hyper plane minimizing training error, the function has to have a form as bellow (Izenman, 2008):

$$d\left(x,w,b\right) = w^T x + b = \sum_{i=1}^{n} w_i x_i + b \quad (6)$$

where, in the above equation, x is the data in hand and b is the bias parameter. The intended hyper plane is derived by solving the general equation bellow (Izenman, 2008):

$$\min\left\{\begin{array}{l}\frac{1}{2}w^2 + C^+\sum_{y_i=1}^{n}\xi_i + C^-\sum_{y_j=-1}^{n_-}\xi_j \\ -\sum_{i=1}^{n}\alpha_i[y_i\left(wx_i - b\right) - 1 + \xi_i] - \sum_{i=1}^{n}\beta_i\xi_i\end{array}\right\} \quad (7)$$

Constraints of this problem are as follows (Izenman, 2008):

$$0 \le \alpha_i \le C^+ if y_i = 1 \quad (8)$$

$$0 \le \alpha_i \le C^- if y_i = -1 \quad (9)$$

where, C^+ and C^- are the cost of mis-categorization of positive and negative classes, respectively. In ordinary SVM these two values are equal. If data is an imbalanced data set, two different values for C is used, which is named cost sensitive Support Vector Machine (SVM-C). Kernel function used in this study is Radial Basis Function. This function has γ parameter which is obtained in training stage along with C parameter from data (Izenman, 2008).

Logit Model

Logit model is considered one of the choice models and is a parametric method. The fundamental basis for various choice models is the maximization of utility behaviour of economics (Hensher *et al*., 2005). Calibration of these models is performed based on maximizing utility of each choice (Train, 2007). According to probability density function assumed for error term ε_{in}, type of the discrete choice model is defined. If the unobserved part is assumed to follow Gumbel distribution, differences between error terms follows

logistic distribution and it gives rise to Binary logit model (Train, 2003). Due to the closed form integral of logit model, the probability of a binary logit model becomes a simple equation as bellow (Train, 2003):

$$P_{n1} = \frac{e^{U_{n1}}}{e^{U_{n1}} + e^{U_{n2}}} \tag{10}$$

$$P_{n2} = \frac{e^{U_{n2}}}{e^{U_{n1}} + e^{U_{n2}}} \tag{11}$$

Above equations can be rewritten as follows (Train, 2003):

$$P_{n1} = \frac{e^{U_2}}{e^{U_2}(1 + e^{U_1} / e^{U_2})} = \frac{1}{1 + e^{U_1 + U_2}} \tag{12}$$

Evaluation Criteria

In current study for dividing data and training network, Cross-Validation Accuracy (CVA) method of *K* is employed. This method is used because of small number of rail observations in data. In this method, first of all, the data set is divided into *K* mutually exclusive subsets of almost the same size (Baykan and Yilmaz, 2011). Each time classification model is trained over all partitions except for one and it is examined in this partition. Cross-validation accuracy is the estimate of the accuracy of the model by simple average of *K* individual accuracy of predictions calculated (Baykan and Yilmaz, 2011):

$$CVA = \frac{1}{K}\sum_{i=1}^{K} A_i \tag{13}$$

where, *CVA* is cross-validation accuracy, *K* is number of partitions and A_i is accuracy of partition *i* (Baykan and Yilmaz, 2011). In order to evaluate introduced algorithms some indices are used. Generally, in condition of a dual classification, irregularities matrix is used to calculate the evaluation indices of models.

According to the data, share of rail mode is less than 10%, if the algorithm of classification, classifies all samples into the truck mode, the percentage of total correct prediction is 83%, while this model is of no use because of the inability to classify data. Therefore in investigating models over imbalanced data, rather than total accuracy, prediction accuracy of each category is also considered. These values can be calculated as follows:

$$predictionaccuracyoftruckmode = nct / (nwt + nct) \tag{14}$$

$$predictionaccuracyofrailmode = ncr / (nwr + ncr) \tag{15}$$

$$totalpredictionaccuracy = (nct + ncr) / (nwt + nct) \tag{16}$$

Where:
Nct = Number of correct truck prediction
nwt = Number of wrong truck prediction
ncr = Number of correct rail prediction
nwr = Number of wrong rail prediction

Kubat and Matwin (1997) suggested G-Means for evaluating models over imbalanced data. This measure can be obtained as follows:

$$G - Means = \sqrt{\begin{matrix}\text{prediction accuracy of truck mode} \\ \times\text{prediction accuracy of rail mode}\end{matrix}} \tag{17}$$

Results

Descriptive Models using C5 Decision Tree

In this study performance of C5 decision tree in context of shipment movement mode choice is discussed from two points of view. That decision trees are powerful in detecting effective variables, first C5 tree is established on the whole data. Therefore all variables in Table 1 are incorporated in model estimation. In this condition, in addition to impedance, truck impedance and rail impedance are used separately. Also commodity type variable is used instead of commodity type binary variable. Effective variables, respective to descriptive strength, are weight, highway impedance and containerized shipment. The point to be mentioned is that some variables of Table 1 are ineffective in forming the tree. Distance variable which in Samimi *et al*. (2011) study is found to be an important variable in this tree is not a good descriptor and therefore not present in the model. Accuracy of the model for truck mode, rail mode and total accuracy are respectively 99.32, 77.78 and 97.70%. Variables' normalized descriptive power are provided in Table 2.

Considering the model proposed by Samimi *et al*. (2012) and results of this model, it can be recognized that effective variables by parametric and non-parametric models produce same outcome.

Prediction Model with C5 Decision Tree

For the sake of examining prediction power of the introduced algorithm in previous section, only significant variables presented by Samimi *et al*. (2012) is used in model calibration. The accuracy is calculated based on cross-validation accuracy. The number of observations excluding inadequate ones is 479, of which 36 observations belong to rail mode choice. In dividing data equally and without a change in rail share in sample, *K* is chosen to be 6. The model is formed in six iterations and two stages, training and testing.

Table 2. Descriptive power of variables in Logit and C5 tree

Model	Logit	Measure	C5	Measure
Variable	Weight	0.562	Weight	0.559
Importance	Distance	0.213	Distance	0.338
	Commodity type	0.105	Containerized	0.086
	Containerized	0.075	Commodity type	0.017
	Impedance	0.042		

Decision tree algorithms are able to fit 100% of data, whence the number of production rules equal the number of observations. In this situation, performance of the resulting tree over unobserved data is too weak. Regarding this issue, restrictions in growth of trees are assumed to prevent the tree from over-fitting training data.

Performance of C5 tree in predicting rail mode excels logit model in all 6 iterations. The total accuracy of these models over both training and testing data is 96.48% for C5 and 95.14% for logit model.

Ranking structure of C5 tree can be seen in Fig. 1. In node 0, called the root node, whole data is present. As it can be seen, first division is conducted based on distance. Samples with distance lower than 804.6 miles are classified in node number 1. About 332 samples, which are 75% of total data belonging to truck mode, are placed in this node. In node 3 according to weight lower than 51,000 pounds rule, 317 observations which mean 71.5% of truck mode samples are classified.

According to the rule, elicited above, firms are inclined to use truck mode for their shipment movement in short distances and also for lightweight freight which is consistent with real situations. 19 observations included in node 3 are also rationally classified according to their respective weight. Again going back to first category node 6, it can be seen that 7 observations which are about 20% of rail mode or classified based on containerized shipment. In node 8, there are 103 observations of truck mode that 38.5% of them are purely classified according to shipment weight lower than 1616 pounds law. This classification based on lower weight appears rational too. Of 25 observations of rail mode in node 10, 26% of them, which is 6 observations, are purely classified according to shipment weight over 75420 pounds rule. This rule also implies that in very high shipment weights and long distances, rail mode is preferred and this seems logical. In node 11, belonging to shipments with weight lower than 78420 pounds, there are 19 observations which is 52.8% of total rail mode choices. About 15 observations in this node are classified according to commodity type. Similarly in node 12, classification is performed based on distance. Node 14 consists of 10 observations of 15 rail mode observations of its upper node according to distance over 1772 miles rule. Also 25 of 29 observations of truck mode in upper node are classified in node 13 based on distances lower than 1772 miles rule. This classification implies that while in short distances truck mode is preferred; in long distances rail mode is given preference.

Prediction Models of Artificial Neural Network and Support Vector Machines

To setup the artificial neural network, a multilayer perceptron (back propagation) with training along with supervision is used. This can be divided into two phases: Propagation and weight update. The two phases are repeated until the performance of the network is good enough. In back propagation algorithms, the output values are compared with the correct answer to compute the value of some predefined error-function. Using this information, the algorithm adjusts the weights of each connection in order to reduce the value of the error function by some small amount. After repeating this process for a sufficiently large number of training cycles, the network will usually converge to some state where the error of the calculations is small. In this case, one would say that the network has learned a certain target function (Dougherty, 1995). The independent variables used in this algorithm are significant variables resulting from the Logit model. As the given solutions are not independent from the defined structure different combinations of these two criterions are examined. The evaluation criterion is the model's correct forecast based on total accuracy and average of G. The accuracy of the model is conducted based on a five part cross-validation accuracy covering only the training data. Each neural network, after being trained over five parts of training, is applied on the sixth part which is the evaluation part and outputs are recorded. The number of layers and neurons in each layer are obtained by the designer during the trial and error process. The final neural network comprises five input neurons respective to each independent variable, a hidden layer of 13 neurons and two neurons respective to the choice of rail or truck. The structure of this network is illustrated in Fig. 2. Input data is normalized between 0 and 1 before training in order to reach better solutions. The training rate is 0.3, the momentum is 0.2 and the maximum iterations for each time of training are restricted to 1000 iterations. The software used in building up the neural network is RapidMiner 4.4 (2009).

In order to train the support vector machine, firstly it is necessary to recognize two parameters: C and γ. As the evaluation procedure of models which is 6 parts validation, models are generated in 6 iterations. In each one of iterations based on training data and by using five parts validation, C and γ are derived by Grid Search Algorithm.

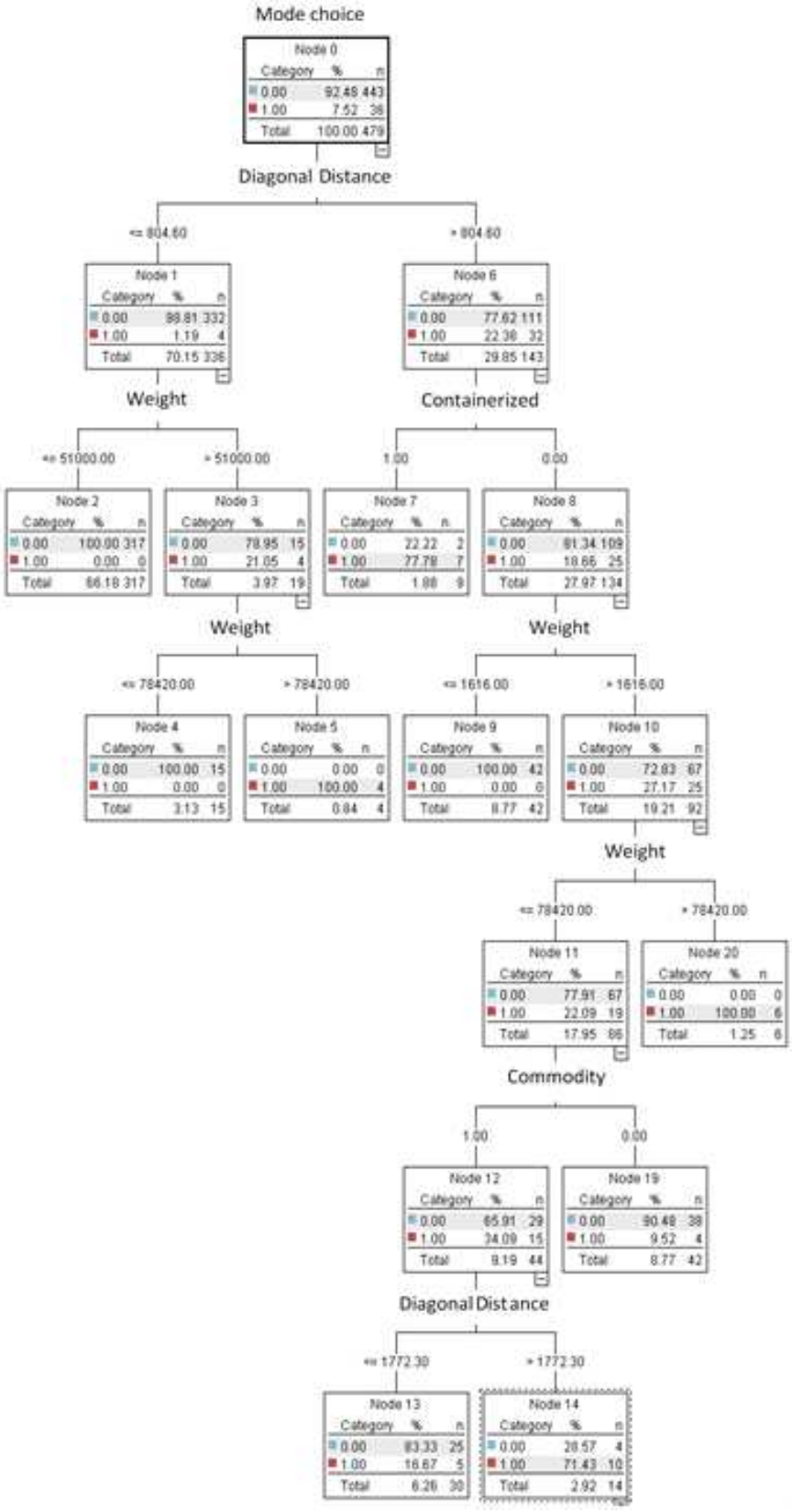

Figure 1. C5 tree of prediction model

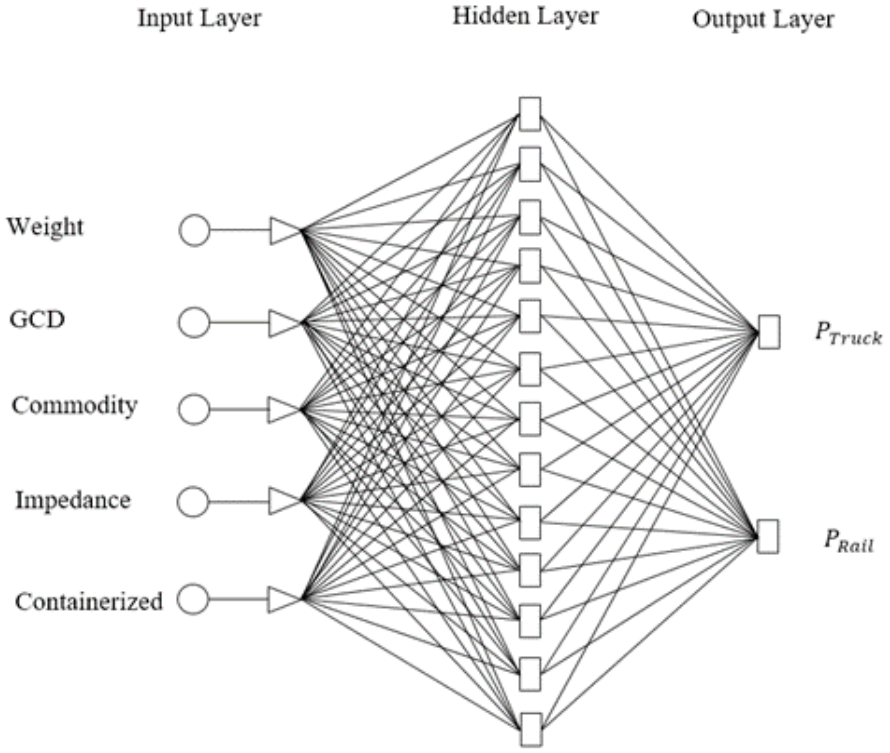

Figure 2. Artificial Neural Network structure

Then in each one of the iterations, the model is examined using the whole training data. In constructing the cost sensitive support vector machine, in each iteration, cost of wrong rail mode categorization is added and then two criterions, total accuracy and average of *G*, are evaluated, so that the optimal condition can be obtained.

Discussion

Comparing Models Introduced in Prediction Accuracy

Now it is time to compare other algorithms, which are logit model, C5 decision tree, artificial neural networks and ordinary and cost-sensitive support vector machines. Table 3 contains prediction accuracy of truck and rail mode for these models.

According to Table 3, in training stage, based on correct prediction of rail samples, the SVM-C model has the best performance; and then, neural network model, C5, SVM and logit produce the best performance respectively. Also according to G index, standing of models is the same as above; which means logit model showed the worst performance. In evaluation stage, the SVM-C model showed the best performance in prediction of rail mode. Precision of models based on this criterion is illustrated in Fig. 3. Based on accuracy of prediction, the next model with the best performance after the SVM-C model is the C5 tree model. Other models, in evaluation stage, showed mostly identical performance.

Studying the Explanatory Power of Variables in Decision Trees

In order to recognize the explanatory power of each variable, Logit model and C5 decision tree are estimated on the whole data. Recognizing the variables which play the most fundamental role in predicting the dependent variable, is one the most important parts in modelling. Table 3 illustrates the descriptive power of each variable based on introduced measures defined in each method. The index delineates proportion of explaining the dependent variable over the whole data. In computing the measure of explanation of each variable in Logit model, the difference between McFadden indices of a model with complete set of variables and a model without the respected variable has been used (Train, 2003). Mentioned quantities are normalized to one so that they will be comparable to those of the decision tree. In each method, the summation of performed measures equals to one which results in representing proportion of explanation of each variable of the whole.

Table 3. Mean prediction accuracy of Logit, SVM, Neural network, C5 tree and SVM-C

		Correct predictions in training stage(%)				Correct predictions in examining stage(%)			
Model	6 iterations	Truck	Rail	Total	G-Index	Truck	Rail	Total	G-Index
LOGIT	Mean	98.5	55.6	95.2	73.8	98.0	58.3	95.0	74.9
	Standard deviation	0.60	4.04	0.72	2.42	2.53	17.48	2.95	11.92
SVM	Mean	99.6	56.7	96.3	74.9	99.6	55.6	96.2	74.2
	Standard deviation	0.14	9.19	0.64	6.11	0.70	8.61	0.78	5.57
NEURAL	Mean	99.0	76.1	97.3	86.7	96.6	61.1	93.9	75.7
	Standard deviation	0.78	6.80	0.49	3.63	1.44	22.77	1.48	13.71
C5	Mean	99.4	75.6	97.6	86.6	97.3	66.7	95.2	80.1
	Standard deviation	0.37	4.04	0.49	4.10	2.26	14.91	2.15	9.02
SVM-C	Mean	97.7	78.3	96.2	87.4	95.9	72.2	94.2	82.8
	Standard deviation	0.70	5.87	0.56	3.15	1.48	17.21	2.04	9.97

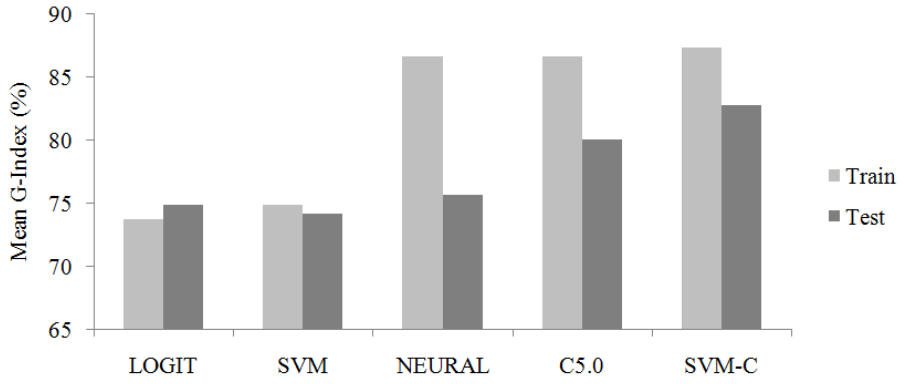

Fig. 3. Comparing performance of models based on mean G-Index

Referring to Table 2, in C5 model except resistance index the rest of variables entered logically the classification process. By comparing measures of C5 and logit model it can be inferred that significance of variables in explaining target variable is different. The only variable with equal effect is weight. The effect of distance in C5 model is approximately 12% more than logit model. Also the importance of container in C5 model is greater than type of good while it is the contrary in logit. The index of resistance has not entered classification tree at all.

Study the Effect of Omitting Variables on the Power of Prediction of Models

The aim in this part is to study the extent to which, model's prediction accuracy is dependent on independent variables in logit models, neural network, ordinary support vector machine and cost sensitive support vector machine. Therefore above models are estimated while omitting one variable. The predictions of accuracy are obtained in six iterations. To evaluate the effect of omitting each variable, reduction in G index is calculated. Figure 4 illustrates the reduction in G index for each model in both training and examining stage. The weight variable is the most effective variable in all four models. The maximum reduction in G index is caused by omitting this variable from SVM model, which is approximately 45% of examination data. After SVM, logit and SVM-C models experience the most reduction in prediction accuracy after omitting this variable. Artificial neural network has the least reduction in prediction accuracy that is about 27%. Another important variable is distance which results in 15% reduction in prediction accuracy in logit model. After logit model, the most reduction in prediction accuracy can be seen in SVM-C, logit, ANN and SVM models.

It can be seen generally that neural network model in examination stage has shown the least reduction in prediction accuracy after omitting each one of variables. The next model to be mentioned beside neural network model is SVM-C. Except for weight variable, maximum reduction in prediction accuracy, caused by omitting each variable, belongs to logit model. This result shows that logit model outputs are more dependent on variables than data mining models while other models excel in predicting target variable if there is lack of sufficient data on hand. It can be mentioned that in C5 tree, omitting variables such as resistance index or container has no effect on results, which is because of using some variables for prediction.

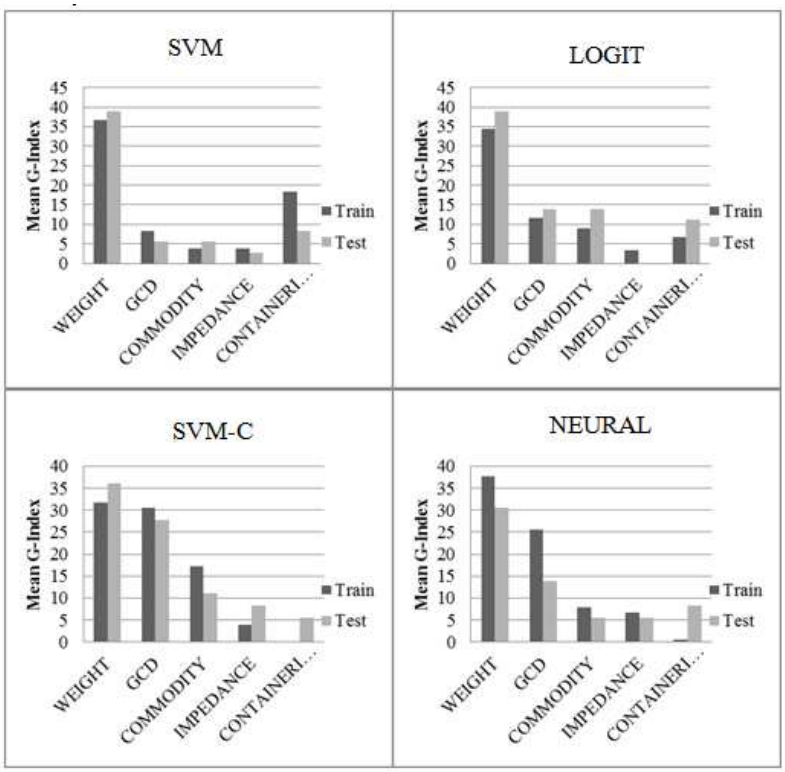

Fig. 4. Effect of omitting variables on mean G-Index in models

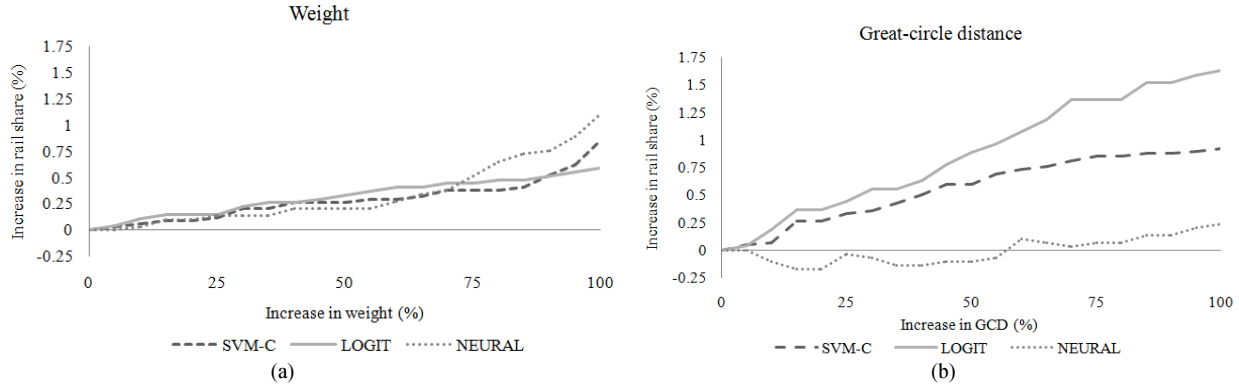

Fig. 5. (a) Rail share change respected to change in shipment weight, (b) Rail share change respected to change in shipment GCD

Sensitivity Analysis of Models

In this stage neural network and cost sensitive support vector machine along with logit model are to be used. The reason of choosing these models is their appropriate performance in prediction. Two important parameters, namely distance and cost, are chosen for sensitivity analysis. First three models are estimated on whole data. In the next stage weight variable is increased in steps of 5% and then rail share is recorded. This procedure is also done for distance variable. Figure 5 illustrates the trend of change in rail share as these two variables increase. As it can be seen from the figures, for weight variable, the two data mining models represented the same behaviour as logit model. Generally because of robust statistic fundamental of logit model, it is the base for comparing performance of machine learning algorithms. Artificial neural network does not represent tangible behaviour for distance variable; even in some points, reduction in rail share is shown while logit model shows a positive relation between distance and rail share. SVM-C model represented rather appropriate performance. A noteworthy point is that, it is expected that these models represent more logical prediction in

low increase amounts because by increasing amounts, their change interval exceeds the maximum amount of respected variables in training sample. Models perform weaker out of their training data.

Conclusion

This study aims to investigate the performance of various data mining models in freight mode choice. To compare them, logit model, the most prevalent model in this context is employed. It is established on robust statistics basis; therefore results obtained from this model are attributable and may be viewed as comparison base. To evaluate models cross-validation accuracy method is used. The fundamental criterion for evaluating performance of models in prediction is accuracy in predicting rail mode as it constitutes a small portion of data. G-Index is another measure for evaluating models.

According to structure analysis of C5 tree, long distance between origin and destination, bulk shipments, containerized shipment and the shipment being a type of agricultural, chemical, pharmaceutical, gravel or sand, cement, metal, compound commodity or prerequisite foodstuffs, increase the chance of choosing rail mode.

The most notable benefit of decision trees is their graphical output. These models offering if-then rules in sequential order along with straightforward interpretation of relations between descriptive variables and dependent variable are able to classify unobserved data with adequate accuracy. One of their major problems is a lack of confidence interval for dividers in each node. A characteristic of these trees which restricts the use of them is complicated structure of the tree if the output class is dependent on too many variables. In such a situation, determining optimal parameters to form the tree also becomes arduous. In addition, the final tree will have multiple nodes making their exhibition and interpretation tough too (Izenman, 2008).

After evaluating C5 tree, prediction models using artificial neural network and ordinary and cost sensitive support vector machines were generated. According to results in both training and testing stage, SVM-C model gives the highest fit.

Then sensitivity of prediction accuracy of the models by omitting each variable from data was examined. Results indicated weight as the most important variable. Artificial neural network and SVM-C experienced the least decrease in accuracy. At last, sensitivity analysis was conducted on artificial neural network and SVM-C and logit model as comparison base, on variables distance and weight. In this stage SVM-C exhibited more logical behaviour confronting changing weight and distance measures.

In general, it can be argued that one of the most important benefits of using data mining methods is that in modelling process, there is no need to make statistics assumptions. The most important statistical assumption encountering researchers with unacceptable answers is multi-collinearity between independent variables while none of the models introduced in this study have such problem. Besides suitable performance, establishing these models does not require much skill. Generally, there is no need to define model structure in these methods. Perhaps, because of model misspecification, statistics models may give rise to wrong answers. The main deficiency of these algorithms is their inability to capture marginal effects or elasticity for effective variables unlike choice models. The values obtained in this study provide valuable results for interpreting the effect of various parameters and policy making in this context.

The methodology presented in this study may be a suitable pattern for evaluating performance of data mining models in other issues of transportation. Also in this study, for the first time in modelling history of freight movement mode choice, various algorithms are used. Due to the dependency of results of the models on data, testing these models over other databases is suggested.

Funding Information

The authors have no support or funding to report.

Author's Contributions

Amir Samimi: His main contribution is in research design, data collection and reviewing the manuscript.

Hesamoddin Razi-Ardakani: His main contribution is in research design, mathematical concept and writing the manuscript.

Amir Nohekhan: His main contribution is in research design, mathematical concept and writing the manuscript.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

Abdelwahab, W. and T. Sayed, 1999. Freight mode choice models using artificial neural networks. Civil Eng. Environ. Syst., 16: 267-286. DOI: 10.1080/02630259908970267

Baykan, N.A. and N. Yilmaz, 2011. A mineral classification system with multiple artificial neural network using k-fold cross validation. Math. Comput. Applic., 16: 22-30. DOI: 10.3390/mca16010022

Breiman, L., J.H. Friedman, R.A. Olshen and C.J. Stone, 1984. Classification and Regression Trees. Taylor and Francis, New York, ISBN-10: 0412048418, pp: 368.

Celik, H.M., 2004. Modeling freight distribution using artificial neural networks. J. Transport Geography, 12: 141-148. DOI: 10.1016/j.jtrangeo.2003.12.003

Chang, L.Y. and H.W. Wang, 2006. Analysis of traffic injury severity: An application of non-parametric classification tree techniques. Accident Analysis Prevent., 38: 1019-1027. DOI: 10.1016/j.aap.2006.04.009

Combes, F. and L.A. Tavasszy, 2016. Inventory theory, mode choice and network structure in freight transport. Eur. J. Transport Infrastructure Res., 16: 38-52.

de Dios Ortúzar, J. and L.G. Willumsen, 2001. Modelling Transport. 3rd Edn., Wiley, Chichester, ISBN-10: 0471861103, pp: 499.

Dougherty, M., 1995. A review of neural networks applied to transport. Trans. Res. Part C Emerg. Technol., 3: 247-260. DOI: 10.1016/0968-090X(95)00009-8

Hensher, D.A., J.M. Rose and W.H. Greene, 2005. Applied Choice Analysis: A Primer. 1st Edn., Cambridge University Press, Cambridge, ISBN-10: 0521844266, pp: 717.

Hensher, D.A. and T.T. Ton, 2000. A comparison of the predictive potential of artificial neural networks and nested logit models for commuter mode choice. Trans. Res Part E: Logistics Trans. Rev., 36: 155-172. DOI: 10.1016/S1366-5545(99)00030-7

Izenman, A.J., 2008. Modern Multivariate Statistical Techniques: Regression, Classification and Manifold Learning. 1st Edn., Springer Science and Business Media, Berlin, ISBN-10: 0387781897, pp: 733.

Kass, G.V., 1980. An exploratory technique for investigating large quantities of categorical data. Applied Statist., 29: 119-127. DOI: 10.2307/2986296

Kotsiantis, S.B., 2007. Supervised machine learning: A review of classification techniques. Informatica, 31: 249-268.

Kubat, M. and S. Matwin, 1997. Addressing the curse of imbalanced training sets: One sided selection. Proceedings of the 14th International Conference on Machine Learning.

Loh, W.Y. and Y.S. Shin, 1997. Split selection methods for classification trees. Stat. Sinica, 7: 815-840.

Lu, Y. and K. Kawamura, 2010. Data-mining approach to work trip mode choice analysis in Chicago, Illinois, area. Trans. Res. Record: J. Trans. Res. Board, 2156: 73-80. DOI: 10.3141/2156-09

Moons, E., G. Wets and M. Aerts, 2007. Nonlinear models for determining mode choice: Accuracy is not always the optimal goal. Proceedings of the Aritficial Intelligence 13th Portuguese Conference on Progress in Artificial Intelligence, Dec. 03-07, Guimarães, Portugal, pp: 183-194. DOI: 10.1007/978-3-540-77002-2_16

Nijkamp, P., A. Reggiani and W.F. Tsang, 2004. Comparative modelling of interregional transport flows: Applications to multimodal European freight transport. Eur. J. Operat. Res., 155: 584-602. DOI: 10.1016/j.ejor.2003.08.007

Pal, S.K. and S. Mitra, 1992, Multilayer perceptron, fuzzy sets and classification. IEEE Trans. Neural Netw., 3: 683-697. DOI: 10.1109/72.159058

Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. 1st Edn., Morgan Kaufmann, San Mateo, ISBN-10: 1558602399, pp: 302.

RapidMiner 4.4, 2009. User Guide Operator Reference Developer Tutorial, Rapid-I GmbH, Stockumer Str. 475, 44227 Dortmund, Germany.

Rashidi, T.H. and A. Mohammadian, 2011. Household travel attributes transferability analysis: Application of a hierarchical rule based approach. Transportation, 38: 697-714. DOI: 10.1007/s11116-011-9339-8

Samimi, A., K. Kawamura and A. Mohammadian, 2011. A behavioral analysis of freight mode choice decisions. Trans. Plann. Technol., 34: 857-869. DOI: 10.1080/03081060.2011.600092

Samimi, A., A.K. Mohammadian and K. Kawamura, 2010. Online freight shipment survey in the United States: Lessons learned and nonresponse bias analysis. Proceedings of the 89th Annual Transportation Research Board Meeting, (RBM' 10), Washington, D.C. pp: 15-15.

Samimi, A., Z. Pourabdollahi, A.K. Mohammadian and K. Kawamura, 2012. An activity-based freight mode choice microsimulation model. Int. J. Trans. Res., 6: 142-151. DOI: 10.1179/1942787514Y.0000000021

Sayed, T. and A. Razavi, 2000. Comparison of neural and conventional Approaches to mode choice analysis. J. Comput. Civil Eng., 14: 23-30. DOI: 10.1061/(ASCE)0887-3801(2000)14:1(23), 23-30

Tortum, A., N. Yayla and M. Gökdağ, 2009. The modeling of mode choices of intercity freight transportation with the artificial neural networks and adaptive neuro-fuzzy inference system. Expert Systems Applic., 36: 6199-6217. DOI: 10.1016/j.eswa.2008.07.032

Train, K., 2003. Discrete Choice Methods With Simulation. 1st Edn., Cambridge University Press, Cambridge, ISBN-10: 0521816963, pp: 342.

Train, K.E., 2007. Discrete Choice Methods with Simulation. 2nd Edn., Cambridge University Press, United States of America.

Vapnik, V., 1995. The Nature of Statistical Learning Theory. 2nd Edn., Springer, New York, ISBN-10: 0387945598, pp: 188.

Wisetjindawat, W., H. Oiwa and M. Fujita, 2015. Rare Mode Choice in Freight Transport: Modal Shift from Road to Rail. J. Eastern Asia Society Trans. Stud., 11: 774-787. DOI: 10.11175/easts.11.774

Xie, C., J. Lu and E. Parkany, 2003. Work Travel Mode Choice Modeling with Data Mining: Decision Trees and Neural Networks. Trans. Res. Record, 1854: 50-61. DOI: 10.3141/1854-06

Zhang, Y. and Y. Xie, 2008. Travel mode choice modeling with support vector machines. Trans. Res. Record J. Trans. Res. Record, 2076: 141-150. DOI: 10.3141/2076-16