Extraction of attribute importance from satisfaction surveys with data mining techniques: a comparison between neural networks and decision trees

By: Juan de Oña, Rocío de Oña and Concepción Garrido

Juan de Oña, Rocío de Oña and Concepción Garrido (2017) *Extraction of attribute importance from satisfaction surveys with data mining techniques: a comparison between neural networks and decision trees*. **Transportatio Letters, 9(1), 39-48** DOI: 10.1080/19427867.2015.1136917

Extraction of attribute importance from satisfaction surveys with data mining techniques: a comparison between neural networks and decision trees

Abstract

When a public transport manager conducts a customer satisfaction survey (CSS), the goal is to determine the overall satisfaction of passengers with the service, as well as their satisfaction with specific aspects (e.g., frequency, speed, and comfort). Another fundamental objective is to assess the importance to customers of each attribute individually. Asking directly about this importance involves a number of drawbacks; therefore, most studies extract this importance from surveys that ask questions only about global satisfaction and specific satisfaction regarding each attribute. This paper investigates the capability and performance of two emerging data mining methods, namely, decision trees and neural networks, for extracting the importance of attributes from CSS. A total of 858 surveys about the metropolitan bus service in Granada (Spain) were used to model estimation and evaluation. The main advantages and disadvantages of each method are studied from the standpoint of public transport managers.

Keywords: Service quality; public transportation; artificial neural networks; decision trees.

1. Introduction

When an existing level of service of public transport (PT) cannot compete with the automobile, the effectiveness of transport policies in reducing the use of cars is limited (Beale and Bonsall, 2007; Brög et al., 2009). In this context, PT managers need a tool

for measuring the quality of service delivered so that they can formulate profitable strategies that improve the levels of service in harmony with passengers' requirements (de Oña and de Oña, in press). Operating companies should not only determine the degree of satisfaction about a series of attributes characterizing the service—they should also identify which attributes have the most influence on customers' global assessment of service. This is probably the most important aspect: to know which attributes have the greatest influence on overall satisfaction.

Perceptions are usually measured by means of customer satisfaction surveys (CSS) developed every year, or sometimes every six months, by PT companies. The importance of each attribute is rated by passengers in the survey or is derived by statistically testing the strength of the relationship between individual attributes and global satisfaction. The first approach has several drawbacks (Weinstein, 2000). The survey is longer because each attribute has to be addressed twice: once for perception and once for importance. This means that the number of attributes mentioned in the survey is reduced to save time (PT users have a limited amount of time for full face-to-face surveys). Moreover, there may be insufficient differentiation in importance ratings, with customers rating most items near the top of the scale, or attributes may be rated as important even though they have little influence on overall satisfaction.

For these reasons, there is a growing tendency to extract those attributes that have the most impact on users' global evaluation by derived importance methods. Several methodologies have been used to tackle this issue (de Oña and de Oña, 2015), but in recent years two data mining (DM) techniques have emerged in the context of transit satisfaction. These techniques have produced powerful results in empirical applications in this field. DM techniques overcome some weaknesses or assumptions underlying more traditional models—normal data, linear relationships between dependent and

independent variables, low multi-colinearity, and so on. According to Garver (2003), these assumptions are almost always violated in customer satisfaction research.

Decision trees (DTs) and artificial neural networks (ANNs) have already been used to analyze user perceptions of different PT services. De Oña et al. (2012) adopted DTs to identify the key factors affecting satisfaction with a bus service operating in Granada (Spain), and subsequently de Oña et al. (2014) and de Oña et al. (2015) applied DTs to a rail service in the North of Italy. Garrido et al. (2014) investigated the most influential factors affecting users' overall satisfaction about the same bus service operating in Granada by adopting ANNs. Along the same line of research, the present study focuses on finding out which of the two DM techniques is more suitable for analyzing service quality from the point of view of PT managers. In this paper, both methodologies are applied to the same service, in order to extract the advantages and disadvantages of each (complexity of the model, time required, difficulty in interpreting the results, fitting parameters, identification of key factors, etc.) and to provide future guidelines for transit satisfaction evaluation by means of DM techniques.

This paper is structured as follows. The next section outlines the data used and briefly describes the users' characteristics and opinions about the service. This is followed by an explanation of DT and ANN methodologies. The outcomes obtained with each methodology are then presented and the results are compared. Finally, the main conclusions of this study are summarized.

2. Data

The data used to implement the DM models came from a CSS conducted by the Transport Consortium of Granada in March 2007. It was a non-research-oriented

survey, involving a rather simple statistical frequency analysis. The 858 respondents were randomly sampled in face-to-face interviews at the main stops of the metropolitan public bus service of Granada. Granada is a medium-sized city in southern Spain, whose metropolitan PT system carried more than 10 million passengers in 2007. The bus system consists of a radial network with two entrances to downtown Granada, one via the north and the other via the south, while 15 bus companies connect the urban agglomerations of the metropolitan area. This structure is due to the fact that over 80% of trips are between the metropolitan boroughs with Granada municipality.

The whole survey database consists of five data sets reflecting passengers' demographic profile, travel behavior, importance of service attributes, perceived service quality attributes, and global evaluation of service quality. Table 1 is a brief summary of the passengers' demographic profile and travel behavior, while Table 2 displays the average and standard deviation rates for the importance of service attributes, perceptions of service attributes, and global satisfaction. Twelve attributes were used to evaluate the service, and a numeric 11-point scale (from 0 to 10) was used for the importance and perception ratings.

(Table 1 here)

Regarding passengers' demographic profile and travel behavior (Table 1), most respondents were female (67%). More than half were aged 18–30 (56.5%) and only 9.5% were older than 60. The majority (61.1%) owned a private vehicle. Roughly half of the respondents indicated that their trip was related to work (29.4%) or study (22.9%). The rest (47.7%) traveled for other reasons. Respondents were asked how frequently they traveled on the bus system per month. The vast majority (88.5%) traveled almost daily or very frequently, while a few reported traveling occasionally or

sporadically (11.5%). The most usual complementary mode used for reaching the bus stop or for reaching the final destination from the bus stop was on foot (77.6% and 94.5%, respectively). Other complementary modes constituted a very small percentage. Finally, the consortium card and the standard ticket were the most widely used types of tickets among passengers, together representing 90.8% of the sample.

(Table 2 here)

Judgments about the importance of the attributes show that the average value of the importance rates is concentrated at the top of the scale (between 8.62 and 9.14). Therefore, this importance is uniform and practically equal in all the attributes. This is one of the serious drawbacks encountered when studying the importance of variables based on the stated opinions of passengers (de Oña et al., 2012; Weinstein, 2000). Moreover, there are similar and low values of the standard deviation (s.d.) among the attributes (<1.82); therefore, their opinions are quite homogeneous.

In contrast, judgments regarding perceptions show greater differences among attributes. They are concentrated in a range from 6 to 8, and users' perceptions are more heterogeneous, with values of s.d. higher than those obtained for the importance rates (from 1.82 to 2.56). The attribute judged as the most heterogeneous is Fare, which is also the attribute with the lowest average rate (6.44). This low perception rate does not necessarily mean that users are dissatisfied with the fare; it could be that users believe a good evaluation of Fare might encourage the PT company to increase the price of the ticket. Nonetheless, the values of attribute perceptions are quite good—all the attributes are perceived to have at least an adequate quality (>6) and some quite a good quality (>7). The attributes characterized by the highest levels of quality were Driver Courtesy, Safety on Board, and Bus Interior Cleanliness.

The overall satisfaction shows an average rate of 7.10. This means that passengers are quite satisfied with the service, and this evaluation is also quite uniform among passengers (s.d. = 1.60).

3. Methodology

3.1. Decision trees

DTs constitute a DM technique used for the classification and prediction of a target variable. Depending on the nature of the variable, two different types of DT models can be developed: if the target variable is discrete, a classification tree is built and the outcome to be predicted is a discrete class, whereas if the target variable is continuous, a regression tree is generated and a numeric quantity is predicted.

There are many different algorithms to generate these models. The main difference among them lies in the partition criterion used for the tree growth. The development of the DT is characterized by the definition of the following steps (Montella et al., 2012): (a) the partitioning criterion to define the optimality function when choosing the best partition of the objects into homogeneous subgroups; (b) the stopping rule to halt the growth of the tree; and (c) the assignment rule to identify either a class or a value as a label of each terminal node. In the following, we focus on the framework of the CART algorithm (Breiman el al., 1984) in view of the good results reached in previous work (de Oña et al, 2012; in press, 2014) using this algorithm for similar purposes. Moreover, a regression tree is applied because this study aims to predict the expected evaluation of satisfaction perceived by an individual as a continuous variable (on an 11-point numeric scale).

Figure 1 shows the steps required for training a DT model and calculating the importance of the predictor variables. The database is randomly divided into M subsets, each containing $(M-1)/M$ portions of the sample (step T01). Common values for M are 5 or 10 (Witten and Frank, 2005). A DT is built for the first subset m (T02), using the group of data $(M-1)/M$ as the training sample and the remaining group of data $1/M$ as the test sample. This is the well known m-fold cross-validation technique (Witten and Frank, 2005). The tree model is developed by using variables i as predictors (these variables are the I attributes that characterize the PT service) and the following considerations (T03):

(a) The partitioning criterion used for evaluating the set of candidate splitting rules is based on the least square (LS) error criterion. Seeing the LS function as an impurity measure of a node, the "worth" of a split will be evaluated by the reduction achieved in the impurity of the parent node in terms of the LS criterion. CART performs all possible splits on each of the independent variables, and the one that best reduces impurity in the parent node is selected. This impurity can be measured as follows (Yohannes and Webb, 1999):

$$\mathrm{Err(t)} = \frac{1}{N_t}\sum_{i=1}^{N_t}\left(y_{i(t)} - \bar{y}_{(t)}\right)^2 \qquad (1)$$

where Err(t) is the impurity function at node t, $y_i(t)$ are the individual values of the independent variable at node t, $\bar{y}_{(t)}$ is the mean value of the target variable at node t, and N_t is the number of instances at node t.

(b) Two stopping rules are applied to the growing procedure:

(b.1) the best splitting criterion among the possible splitters is no greater than 0.0001;

(b.2) the number of cases in one or more child nodes is less than 1% of the whole sample.

(c) The assignment rule used to impute a value, as a label of each terminal node, is the mean value of the target variable at the terminal node.

(Figure 1 here)

The variance of the data explained by the model is calculated on the test sample (T04). Thus, the explained variance by the model will be obtained from the mean square error across the terminal nodes of the built model. Then, the improvement that a variable *i* produces when it is used as the main splitter or substitute splitter is added across each partition of the DT and weighted by the number of cases affected by this improvement (T06). The importance value of the variable *i* is stored (T07). This procedure is repeated from T06 to T07 until *i* reaches *I* (T11), and the importance of the *I* variables for the subset *m* is stored (T10).

Next, the procedure from T03 to T10 is repeated again until *m* reaches *M* (T12). At this point, the predictive accuracy of the DT model (T13) is calculated as the mean value of the variance explained at each of the *M* models stored in step T04. Likewise, the average importance and standard deviation of each variable is calculated (T14) from the *M* values of importance stored for each variable at T10. The ranking of relative importance of each variable is determined by following the criterion that the higher the average value for each variable, the greater its relative importance in the global ranking (T15).

3.2. Artificial neural networks

To calculate the relative importance of the variable under study, several authors have highlighted the advantages of working with ANN sets instead of using a single ANN (e.g., Garrido el al., 2014; Cao and Qiao, 2008, Paliwal and Kumar, 2011; de Oña and Garrido, 2014). We opted to follow this procedure. Figures 2 and 3 show the steps required for training an ANN model and for calculating its accuracy.

The database is randomly divided into three groups—training, validation, and test sets (N01)—and the variables of the whole database are normalized, that is, a range of values in the interval [0,1] are used as input values for every variable, instead of the original interval [0,10] obtained from the surveys, (N02). The ANN typology is the multilayer perceptron (MLP), characterized by being a supervised network. Because many authors (e.g., Funahashi, 1989) have demonstrated that an MLP with one hidden layer is a function universal approximator, we adopted an ANN architecture featuring an input layer with I neurons (as many as there are attributes considered in the study), H neurons in the hidden layer, and J neurons in the output layer (N04).

A collection of synaptic weights connects each neuron with all of the neurons of the following layer. These connections indicate the intensity of the interaction between each pair of neurons (Palmer and Montaño, 2002). Each neuron also has a bias or activation threshold, whose value determines the global potential that must be reached for the neuron to be activated (Martín del Brío and Sanz, 2006). HT different ANN architectures are defined, and each H architecture is trained N times. All of the synaptic weights and biases are randomly initialized with small values to optimize the training performance (N06), and subsequently the MLP training starts by using the gradient descent algorithm, with logarithmic sigmoidal activation functions, and the momentum

and learning rate factors that accelerate the convergence of the training toward a local solution (N07).

Once the training of the ANN architecture with *H* neurons in the hidden layer has finished, the generalization error is determined through the mean absolute percentage error (MAPE) approach (Delen et al., 2006), and this output is stored (N08):

(Figure 2 here)

$$\mathrm{MAPE} = \frac{1}{T} \cdot \sum_{i=1}^{T} \mathrm{abs}\left(\frac{\text{Actual value } i - \text{Set point value } i}{\text{Set point value}} \right) \tag{2}$$

Figure 2 shows that this procedure continues from N06 to N08 until *n* reaches *N* (N10), which indicates that the subset with *H* hidden neurons is complete, and the following subset, with $H = H + 1$ (N11) hidden neurons, must be adequately trained until *HT* ANN architectures are obtained (N12). Each *H* architecture has a different number of hidden neurons.

The average and standard deviation of MAPE are calculated from the *HT* architectures that make up each *H* subset (N13) to select the best ANN subset, that is, the one that provides the lowest values of the average and standard deviation (N14). At this point, we work with the suboptimal ANN subset to determine the relative importance of the study variable. Figure 3 shows detailed flow diagrams for determining the relative importance of the predictor variable, based on two classical methods: connection weights (Olden and Jackson, 2002) and profile (Lek et al., 1995). Both have already been demonstrated to be successful in obtaining a homogeneous hierarchy of importance for the variable under study (Garrido et al., 2014; de Oña and Garrido, 2014).

The connection weight method has been used in various research fields (Gevrey et al., 2003). In this study, this method is applied to each *I* variable contained in each of the *HT* ANNs of the selected subset (N15 in Fig. 2 or N16 in Fig. 3*a*). Hence, for each *I* predictor variable, we obtain *N* results, and we calculate the average (N17). The relative importance of each variable (ranking) follows the criterion that a higher average value signals a greater relative importance of a given variable in the global ranking (N18).

(Figure 3 here)

Figure 3*b* shows the procedure for the profile method (Lek et al., 1995). The profile method is applied for each *I* variable of each of the *HT* ANNs that belong to the suboptimal subset *H* (N19), so that a beam of *N* profiles of variation appears when the profile of variation of each *I* variable is graphically represented (N20), and an average profile of variation is calculated (N21). Then the difference between the maximum and minimum values of the average profile of variation on the ordinate axis is calculated (N22), and finally a ranking of relative importance of the predictor variable is established (N23). A variable having a greater range of difference in the values of the average variation profile is more important than variables whose profiles have a lower range.

A more detailed explanation of the connection weight and profile methods is given by de Oña and Garrido (2014).

4. Results and discussion

Decision trees

The variance according to the DT model is 49.7% of the total variance, indicating that the model accuracy is low (de Oña et al., 2012). Table 3 shows the importance ranking

of the variables according to the normalized rate obtained using the DT method, as well as the average rate for each attribute. This model was calibrated using SPSS software.

(Table 3 here)

The key variables influencing users' overall satisfaction are those more closely related to the operation of the service. The importance rates have been normalized by assigning for each algorithm a normalized value of 100% to the attributes with the highest extracted importance rate, in order to be able to compare the scales and the rankings provided by the DT and ANN algorithms. Hence, the remaining importance rates derived for the individual attributes are referred to the highest one for each algorithm. Thus, Frequency, Speed, and Punctuality show normalized importance values higher than 85%. Temperature, Information, Safety, and Courtesy are also important, with rates around 60% (65.23%, 61.61%, 60.59%, and 59.42%, respectively). The remaining variables have less impact on satisfaction. They are related to Fare, Comfort (Space and Cleanliness), and accessibility of the service (Accessibility and Proximity). In fact, Accessibility is the least important variable with regard to overall satisfaction (21.59%).

A major disadvantage of the DT methodology is that the tree models are very "unstable" (Chang and Wang, 2006). Once the model has made a decision about a variable on which to split the node, the decision cannot be revised or improved, because there is no backtracking technique (Xie et al., 2003). Then, depending on how the sample is stratified, different models may be obtained, and therefore different importance rates could be extracted. On the other hand, however, the results are easily interpreted. They are displayed on graphical charts from which informative "If-Then" rules can be extracted.

Artificial neural networks

Table 4 shows the MAPE average and standard deviation of the 30 sets considered, with 50 ANNs trained in each group. MATLAB software was used to train the ANN architectures. The number of sets and the number of neurons to be trained in each set were chosen according to the criterion followed in previous studies (Garrido et al., 2014; de Oña and Garrido, 2014). Table 4 shows that the accuracy achieved by the trained ANNs is very high—around 95% for all the architectures. Other authors (e.g., Martín del Brío and Sanz, 2006) have underlined that one of the main strengths of ANNs lies in their capacity to find highly nonlinear relationships among study variables, which leads to a high prediction capability, provided that a suitable number of data are available during the training phase. Moreover, studies using MLP ANNs to analyze service quality in other fields (Mahapatra and Khan, 2006; Larasati et al., 2012) have reported very high accuracies.

The set with six neurons in the hidden layer presents the best behavior globally. Given that its average and standard deviation are the lowest, it was selected for the following steps.

(Table 4 here)

So far, the main drawback attributed to ANNs is their inability to determine an homogeneous ranking of relative importance for the variables under study when different methods are applied to the same ANN (Martín del Brío and Sanz, 2006); for this reason, ANNs have been included among the techniques referred to as "black boxes" (Karlaftis and Vlahogianni, 2011). Several authors (Paliwal and Kumar, 2011; Palmer and Montaño, 2002; Olden and Jackson, 2002) have tried to overcome this limitation, with some proposing that the problem can be mitigated by working with sets of ANNs instead of using a single ANN (e.g., 14). Recently, de Oña and Garrido (2014)

and Garrido et al. (2014) have shown that the approach for the MLP described in the section above on "Data" provides homogeneous outcomes for the relative importance of the variables, even when this approach is applied using traditional methods such as connection weights (Olden and Jackson, 2002), perturb (Yao et al., 1998), profile (Lek et al., 1995), and partial derivatives (Dimopoulos et al., 1995). The connection weights and the profile methods showed lower statistical differences for determining the relative importance than the other two methods (Garrido et al., 2014), so both were selected in this study.

Table 5 shows the importance for each variable using the connection weights (CW) and profile (PR) methods, expressed as percentages. As can be seen, the relative importances of variables obtained by the two methods are very similar, and they coincide in identifying the four most important variables: Frequency (100.0% by both PR and CW), Speed (77.7% by PR and 76.0% by CW), Information (64.2% by PR and 66.7% by CW) and Proximity (60.2% by PR and 55.5% by CW). The percentages of relative importance are very similar for the remaining variables, although if a hierarchy is determined by one method, then some variables may change their position with regard to the others. Thus, the results show a level of medium importance in relation to the following six variables: Punctuality (54.5% by PR and 51.3% by CW), Safety (53.3% by PR and 51.4% by CW), Courtesy (48.6% by PR and 47.8% by CW), Temperature (38.4% by PR and 36.6% by CW), Fare (36.4% by PR and 32.0% by CW), and Space (27.2% by PR and 36.5% by CW). Clearly, the two least influential variables by both methods are Cleanliness (3.4% by PR and 27.4% by CW) and Accessibility (17.3% by PR and 14.6% by CW).

(Table 5 here)

Both methods give an approximate idea about the position of every variable in the ranking of relative importance. In addition, both clearly differentiate between the most important and least important variables, and those of medium importance.

Comparison between ANN and DT

The case study CSS database was used to compare the performance of a DT methodology and the two algorithms based on ANN for extracting the importance of attributes from satisfaction surveys in PT.

This research found some differences regarding the performance of these DM techniques. Both ANN algorithms outperformed the DT model in accuracy rates (95% versus 49.7%). This is consistent with the results on accuracy obtained by Xie et al. (2003) and Lee et al. (2010), who compared the performance of ANN and DT models in other fields and arrived at higher accuracy rates for the ANN methodology.

Concerning the importance ranking of the variables obtained with the DT and ANN algorithms (Table 3 and Table 5), both methods provide similar results with respect to the core factors for defining an efficient metropolitan bus public service. These factors were the Frequency and Speed of Operation. Therefore, both should be considered as fundamental in transit planning process and operation/management phases. Other studies support the importance of Frequency (de Oña et al., 2012; Dell´Olio, 2010; 2011; Del Castillo and Benítez, 2013; Tyrinopoulos and Antoniou, 2008).

Conversely, Accessibility and Cleanliness are the characteristics that exert the least influence on users' overall evaluation in both methodologies. Eboli and Mazzulla (2008) also identified Cleanliness as a variable having a low influence on the users of an urban bus service in Cosenza (Italy). Nevertheless, transport companies should not ignore these characteristics, because although they will have almost no influence on

users' overall evaluation when performance quality is high, if performance quality falls, they will probably have a negative influence, leading to a decrease in users' overall satisfaction.

The main differences in the importance rankings of the variables concerned factors with a medium level of importance, which occupied different positions in the ranking, depending on the methodology applied.

If we compare the derived importance rates obtained with both DM methodologies and those stated by the users in the survey (Table 2), some noteworthy differences are seen. Although there is little variation in the importance expressed by passengers, who hold that all the attributes are highly important, in the ranking established for these attributes, Speed occupies seventh position. Yet the data mining techniques deduced Speed as a core factor for the metropolitan service. Likewise, Accessibility and Cleanliness, which were deduced as the variables exerting the least influence on users' overall evaluation, occupy fourth and fifth positions, respectively, in the stated importance ranking. This lack of agreement was also encountered by Weinstein (2000) in a study of the importance of variables based on the stated opinions of passengers.

Table 6 shows a comparison between the advantages and disadvantages of both methods. The main flaw of DT is the instability of the models derived. Depending on the strategy followed for stratifying the sample, the structure and accuracy of the models generated could change, making it difficult to determine the fundamental variables for users. In turn, the main strength of ANN would be its ability to achieve high accuracy in classification and prediction problems. The ranking of importance of the predictive variables is, moreover, stable and consistent. Yet finding the optimal ANN is complicated by the large number of possibilities when choosing the number of neurons

in the hidden layer, the type of activation functions and learning algorithm, or the initial random values selected before training starts. This procedure is wearisome in terms of the time needed to determine the ranking of importance of variables and the time spent choosing the suboptimal ANN set and in the training and testing phases. Several authors have highlighted the complexity of working with ANN (Cao and Qiao, 2008), but, regardless of the ANN chosen, the accuracy of the results is stable (Karlaftis and Vlahogianni, 2011).

(Table 6 here)

Kirby et al. (1997) suggested that accuracy is very important but that it should not be the sole determinant when selecting the proper methodology for prediction; other issues should be considered in selecting the appropriate approach, such as the time and effort required for model development, the skills and expertise required, the transferability of the results, adaptability to changing behaviors, and so on. Likewise, Karlaftis and Vlahogianni (2011) reviewed two different approaches for modeling transportation data, namely, statistics and ANN, and they proposed some guiding questions for transportation researchers to consider when deciding which is the best modeling approach for their analysis—for example, "What are the requirements with respect to accuracy and interpretability of results?" and "How important is interpretability in the problem examined?". Such questions should be used by PT managers and practitioners as a guide for selecting an adequate model for developing a service quality analysis. In choosing between DT and ANN algorithms, to find the option that better addresses their questions, PT managers should weigh up their advantages and disadvantages concerning accuracy, time, interpretability, and expertise required. For example, if an annual routine analysis of the service is performed in order to determine the evolution of the importance ranking of the variables, PT managers could choose a DT approach, whereas

if a detailed analysis is going to be carried out because there is a change in the PT concession, a new public transit service is to be implemented, or significant changes are to be introduced in the service, it could be more appropriate to use an ANN approach that provides greater accuracy.

5. Conclusions

This paper has investigated two emerging DM techniques, namely, DTs and ANNs, in order to determine which is more appropriate for modeling satisfaction in the context of public transportation. Based on data collected with a non-research-oriented survey, very interesting details have been unearthed. Our results serve to confirm the suitability of using this kind of data when advanced modeling techniques are applied, involving collaboration between researchers and industry.

We used 12 predictor variables in this study, but it is possible to work with larger databases and more predictor variables (e.g., by using a large list of attributes describing the service or by including socioeconomic and travel habit variables in order to extract their influence on the model). In such a case, one may wonder whether the advantages of ANNs in terms of accuracy outweigh the disadvantages in terms of time invested to derive the relative importance of factors. Depending on the field of application, it may be preferable to sacrifice accuracy for the sake of speed in calculation and the explanatory capability of DTs.

DT and ANN methodologies share some advantages inherent to DM techniques, such as the ability to discover knowledge in large databases. Furthermore, they are nonparametric models with no underlying model assumptions or predefined relationships between the dependent and independent variables. Both methodologies

exhibit high degrees of flexibility and adaptability of the model structure or parameters to the training data owing to their data induction properties (Xie et al., 2003).

The main disadvantage of ANNs is a matter of explanatory capability, that is, the capacity to determine the relative importance of variables. The procedure used in this study reduces the differences in relative importance considerably, but, even so, the relative importance of variables is not evident and the procedure is tedious. For this reason, the simplicity of the DT model might be preferred by PT managers most of the time (the interpretation of results is facilitated by graphical representation, and they enable the extraction of "If-Then" decision rules, providing explanations for overall satisfaction). Nevertheless, some occasions could require a more precise analysis, and an ANN algorithm might be selected. Accuracy, time invested, interpretability, and expertise required should be considered as determinants for choosing the proper approach that responds to PT managers' and practitioners' questions each specific time.

Moreover, understanding which variables have the greatest influence on users' overall evaluations about the service, together with how they perceive the performance of these variables, helps PT managers to decide which aspects of the service should be improved and how to allocate their resources in the most efficient way according to this information. If these sophisticated methodologies are available for use by PT managers (by programming these methodologies in simple-use software), they would be able to extract interpretative and practical results for formulating specific policy decisions. Additionally, in order to provide further enlightenment regarding users' opinions, advanced sample stratification techniques (e.g., cluster analysis) could be developed in future research to handle users' heterogeneity and to provide better recommendations.

Finally, further research is needed in order to convince PT managers to use these more sophisticated techniques based on data mining approaches instead of other traditional parametric techniques.

Acknowledgements

This study was partially sponsored by the Junta de Andalucía (Spain) through Research Project P08-TEP-03819. The authors also acknowledge Granada's Consorcio de Transportes.

References

Beale, J.R. & Bonsall P.W., (2007). Marketing in the bus industry: A psychological interpretation of some attitudinal and behavioural outcomes. Transportation Research Part F, 10, 271–287.

Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J., (1984). Classification and Regression Trees. Wadsworth, Belmont (CA).

Brög, W., Erl, E., Ker, I., Ryle, J. & Wall, J., (2009). Evaluation of voluntary travel behaviour change: Experiences from three continents. Transport Policy, 16, 281–292.

Cao, M. & Qiao, P., (2008). Neural network committee-based sensitivity analysis strategy for geotechnical engineering problems. Neural Computing and Applications, 17, 509-519.

Chang, L.Y. & Wang, H.W., (2006). Analysis of traffic injury severity: An application of non-parametric classification tree techniques. Accident Analysis and Prevention, 38, 1019–1027.

Delen, D., Sharda, R. & Bessonov, M., (2006). Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. Accident Analysis and Prevention, 38, 434-444.

de Oña, J. & de Oña, R. (2015). Quality of service in public transport based on customer satisfaction surveys: A review and assessment of methodological approaches. Transportation Science, 49(3), 605-622.

de Oña, J., de Oña, R., & Calvo, F. J., (2012). A classification tree approach to identify key factors of transit service quality. Expert Systems with Applications, 39, 11164–11171.

de Oña J, de Oña R, Eboli, L. & Mazzulla, G. (2015). Heterogeneity in perceptions of service quality among groups of railway passengers. International Journal of Sustainable Transportation, 9, 612-626.

de Oña, R., Eboli, L. & Mazzulla, G., (2014). Key factors affecting rail service quality in the northern Italy. A Decision Tree Approach. Transport, 29(1), 75-83.

De Oña, J. & Garrido, C., (2014). Extracting the contribution of independent variables in neural network models: a new approach to handle instability. Neural Computing and Applications, 24(5). DOI: 10.1007/s00521-014-1573-5.

Del Castillo, J.M. & Benitez, F.G., (2013). Determining a public transport satisfaction index from user surveys. Transportmetrica A: Transport Science, 9(8), 713-741.

Dell´Olio, L., Ibeas, A. & Cecín, P., (2010). Modelling user perception of bus transit quality. Transport Policy, 17(6), 388-397.

Dell´Olio, L., Ibeas, A. & Cecín, P., (2011). The quality of service desired by public transport users. Transport Policy, 18(1), 217-227.

Dimopoulos, Y., Bourret P. & Lek, S., (1995). Use of some sensitivity criteria for choosing networks with good generalization ability. Neural Processing Letters, 2, 1-4.

Eboli, L. & Mazzulla, G., (2008). Willingness-to-pay of public transport users for improvement in service quality. European Transport, 38, 107-118.

Funahashi, K.I., (1989). On the approximate realization of continuous mappings by neural networks. Neural Networks, 2, pp. 183-192.

Garrido, C., de Oña, R. & de Oña, J., (2014). Neural networks for analyzing service quality in public transportation. Expert Systems with Applications, 41(15) 6830-6838.

Garver, MS., (2003). Best practices in identifying customer-driven improvement opportunities. Industrial Marketing Management, 32(6), 455–466.

Gevrey, M., Dimopoulos, I. & Lek, S. (2003). Review and comparison of methods to study the contribution of variables in artificial neural network models. Ecological Modelling, 160, 249-264.

Larasati, A., De Yong, C. & Slevitch, L., (2012). The application of neural network and logistics regression models on predicting customer satisfaction in a student-operated restaurant. Procedia Social and Behavioral Sciences, 65, 94-99.

Lee, Ch., Ran, B., Yang, F. & Loh, W.Y., (2010). A hybrid tree approach to modeling alternate route choice behavior with online information. Journal of Intelligent Transportation Systems, 14(4), 209–219.

Lek, S., Beland, A., Dimopoulos, I., Lauga, J. & Moreau, J., (1995). Improved estimation, using neural networks, of the food consumption of fish populations. Marine and Freshwater Research, 46, 1229-1236.

Mahapatra, S.S. & Khan, M.S., (2006). A methodology for evaluation of service quality using neural networks. Presented at the International Conference on Global Manufacturing and Innovation, Rourkela.

Martín del Bío, B. & Sanz Molina, A., (2.006). Neural networks and fuzzy systems. Editorial RA-MA.

Montella, A, Aria, M., D'Ambrosio A. & Mauriello, F., (2012). Analysis of powered two-wheeler crashes in Italy by classification trees and rules discovery. Accident Analysis and Prevention, 49, 58-72.

Olden, J.D. & Jackson, D.A., (2002). Illuminating the "black-box": a randomization approach for understanding variable contributions in artificial neural networks. Ecological Modelling, 154, 135-150.

Paliwal, M. & Kumar, U.A., (2011). Assessing the contribution of variables in feed forward neural network. Applied Soft Computing, 3690-3696.

Palmer, A. & Montaño, J.J., (2002). Neural networks applied to analysis of data. Doctoral Thesis. University of Palma de Mallorca.

Tyrinopoulos, Y. & Antoniou, C., (2008). Public transit user satisfaction: Variability and policy implications. Transport Policy, 15(4), 260–272.

Weinstein, A., (2000). Customer satisfaction among transit riders. How customer rank the relative importance of various service attributes. Transportation Research Record, 1735, 123–132.

Witten, I.H. & Frank, E., (2005). Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufman, Amsterdam.

Xie, C., Lu, J. & Parkany, E., (2003). Work travel mode choice modelling with data mining. In Transportation Research Record: Journal of the Transportation Research Board, 1854, 50-61.

Yao, J., Teng, N., Poh; H.L. & Tan, C.L., (1998). Tan. Forecasting and analysis of marketing data using neural networks. Journal of Information Science and Engineering, 14, 843-862.

Yohannes, Y. & Webb, P., (1999). Classification and Regression Trees. CART: A User Manual for Identifying Indicators of Vulnerability to Famine and Chronic Food Insecurity. International Food Policy Research Institute. Washington D.C.

List of Figures:

List of Tables:

Figure 1.- Flow diagram for training a DT model and calculating the predictors' importance

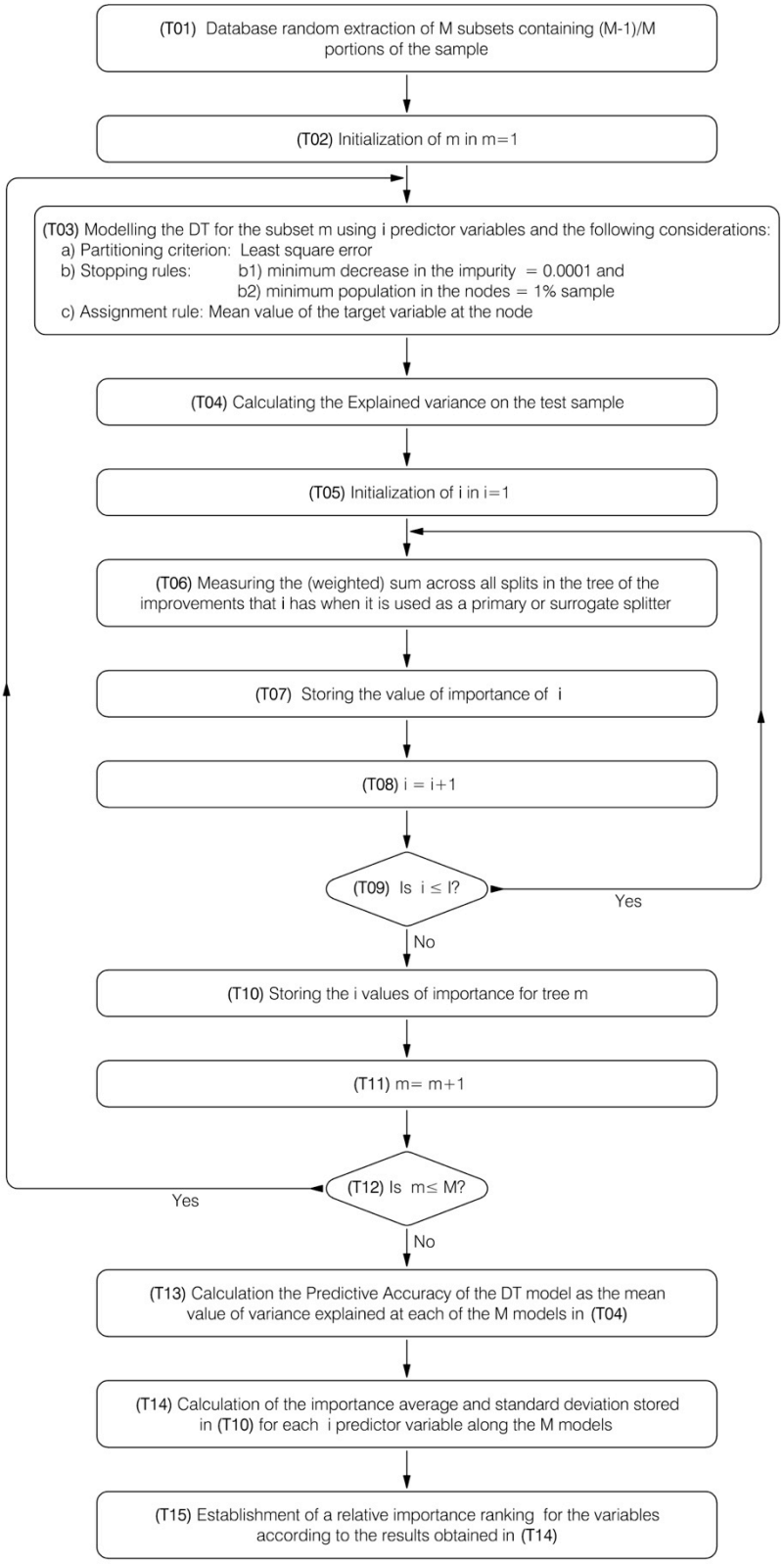

Figure 2.- Flow diagram for training an ANN model

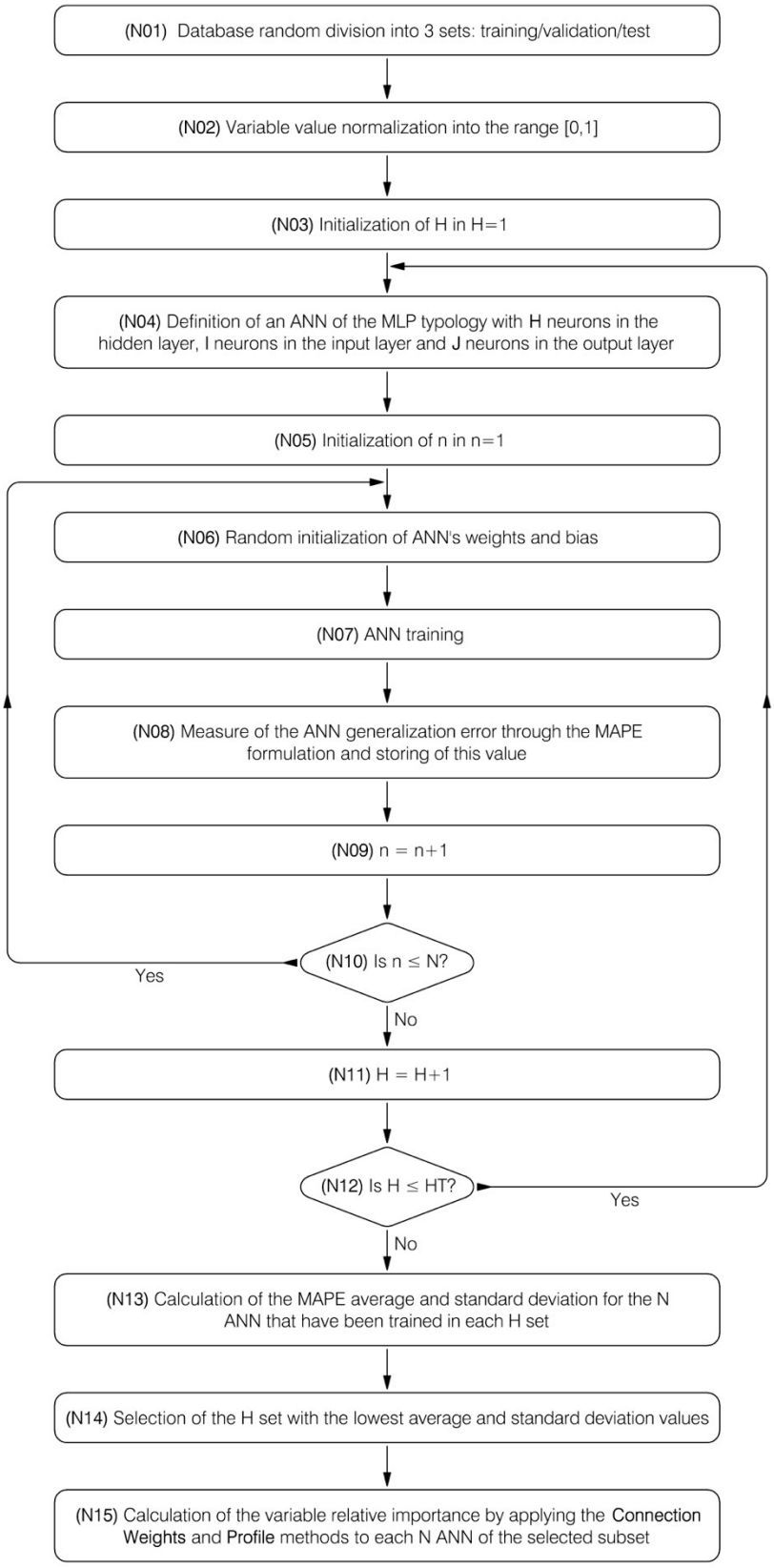

Figure 3.- Procedure for determining variable relative importance by Connection Weights and Profile methods

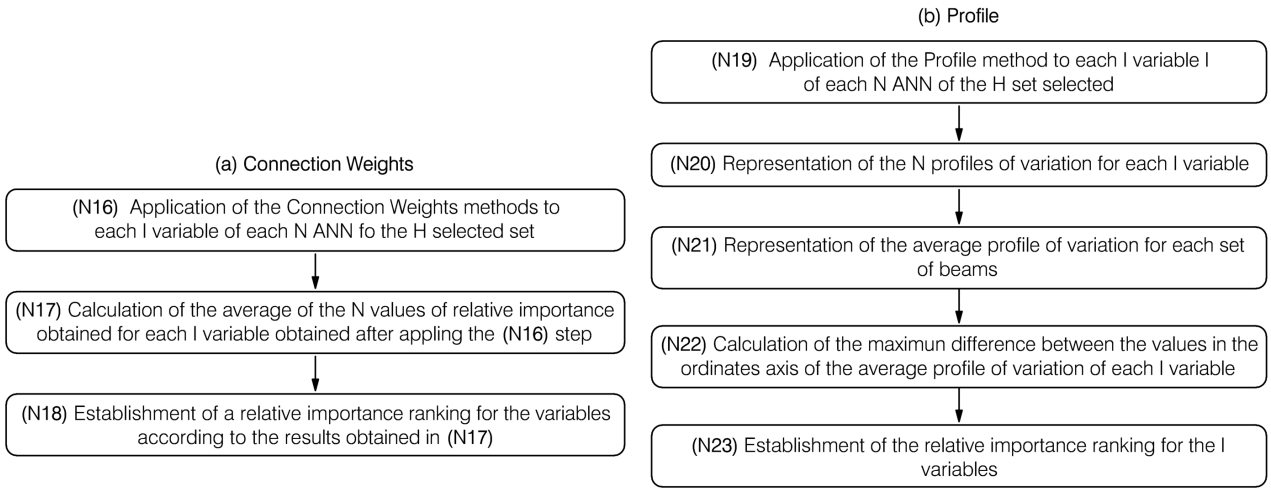

Figure 4.- Ranking of relative importance of each service quality attribute by methods (Connection Weights, Profile, Perturb).

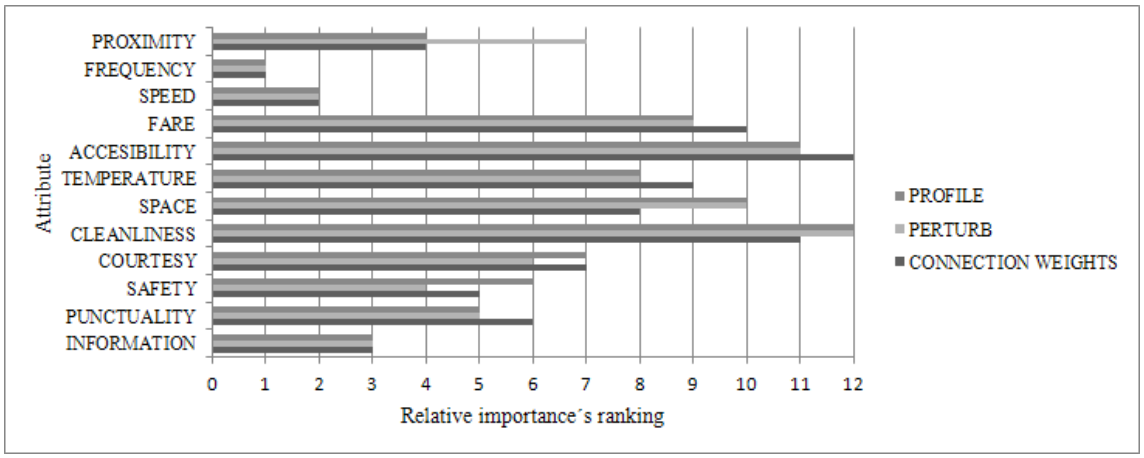

Table 1.- Passengers' demographic profile and travel behavior.

Characteristics	Statistics
1.Gender	Male (33.0%), Female (67.0%)
2.Age	18-30 (56.5%), 31-60 (34.1%), > 60 year-olds (9.5%)
3.Private vehicle availability	Yes (38.9%), No (61.1%)
4.Travel reason	Occupation (29.4%), Studies (22.9%), Doctor (14.2%), Shopping (4.4%), Personal activities (18.7%), Holidays (0.2%), Leisure time (8.6%), Others (1.5%)
5.Frequency	Almost daily (67.9%), Frequently (20.6%), Occasionally (9.0%), Sporadic (2.5%)
6. Complementary modes from origin to bus stop	On foot (77.6%), Car (1.9%), Urban bus (16.9%), Motorbike (0.5%), Others (3.1%)
7. Complementary modes from bus stop to destination	On foot (94.5%), Car (2.1%), Urban bus (2.3%), Motorbike (0.2%), Others (0.9%)
8.Type of ticket	Consortium card (49.6%), Standard ticket (41.2%), Senior citizen pass (4.8%), Others (4.4%)

Table 2. Average values for stated importance and perception rates.

	Importance Rates			Perception Rates	
Attributes	**Ranking**	**Mean**	**Std. Deviation**	**Mean**	**Std. Deviation**
Information	11	8.62	1.73	6.86	2.46
Punctuality	1	9.14	1.45	7.41	2.33
Safety on board	3	8.98	1.53	7.73	1.99
Driver courtesy	6	8.77	1.75	7.96	1.82
Bus interior cleanliness	5	8.86	1.47	7.46	1.84
Bus space	10	8.66	1.72	7.21	2.04
Bus temperature	8	8.72	1.62	7.43	1.97
Accesibility to/from the bus	4	8.91	1.79	6.90	2.48
Fare	6	8.77	1.81	6.44	2.60
Speed	7	8.73	1.71	7.30	1.98
Frequency of service	2	9.05	1.55	6.99	2.56
Proximity to/from origin/destination	9	8.71	1.78	7.43	2.21
Overall Satisfaction				7.10	1.60

Table 3. Ranking of the variables according to the Normalized Importance extracted from the DT approach

	DT	
VARIABLE	**Normalized rate**	**Ranking**
Information (INF)	61.6	5
Punctuality (PUN)	86.3	3
Safety (SAF)	60.6	6
Courtesy (COU)	59.4	7
Cleanliness (CLE)	38.1	11
Space (SPA)	45.8	8
Temperature (TEM)	65.2	4
Accessibility (ACC)	21.6	12
Fare (FAR)	42.6	9
Speed (SPE)	86.8	2
Frequency (FRE)	100.0	1
Proximity (PRO)	41.6	10

Table 4. Average and Standard Deviation values of MAPE for each H ANN architecture

H	Average	Standard Deviation
1	0,053130	0,008159
2	0,053196	0,009032
3	0,052462	0,004534
4	0,051909	0,006709
5	0,052800	0,006206
6	0,049470	0,003578
7	0,051413	0,005187
8	0,051963	0,004378
9	0,051299	0,003613
10	0,053005	0,007362
11	0,050483	0,005071
12	0,051986	0,008620
13	0,052611	0,010607
14	0,051654	0,004267
15	0,052632	0,006461
16	0,051428	0,008729
17	0,051302	0,005589
18	0,051255	0,007039
19	0,051255	0,004630
20	0,051587	0,004929
21	0,050769	0,005591
22	0,052380	0,007812
23	0,051843	0,007030
24	0,049650	0,004700
25	0,050813	0,006672
26	0,052427	0,006302
27	0,051951	0,007792
28	0,050657	0,005582
29	0,053412	0,007483
30	0,051670	0,009091

Table 5. Ranking of the variables according to the Normalized Importance extracted from the ANN algorithms

	PROFILE (PR)		CONNECTION WEIGHTS (CW)	
VARIABLE	**Normalized rate**	**Ranking**	**Normalized rate**	**Ranking**
Information (INF)	64.2	3	66.7	3
Punctuality (PUN)	54.5	5	51.3	6
Safety (SAF)	53.3	6	51.4	5
Courtesy (COU)	48.6	7	47.8	7
Cleanliness (CLE)	3.4	12	27.4	11
Space (SPA)	27.2	10	36.5	9
Temperature (TEM)	38.4	8	36.6	8
Accessibility (ACC)	17.3	11	14.6	12
Fare (FAR)	36.4	9	32.0	10
Speed (SPE)	77.7	2	76.0	2
Frequency (FRE)	100.0	1	100.0	1
Proximity (PRO)	60.2	4	55.5	4

Table 6. Comparison of advantages and disadvantages between DT and ANN

	DT	ANN
Advantages	- Lower complexity for calculating importance rates - Minor time required for determining the relative importance of the variables (seconds) - Model simplicity - Interpretative results because of the graphic representation - It extracts informative "If-Then" rules - The method is not affected by the relationships of the study variables.	- Higher accurary rates - Higher stability for determining the relative importance of the variables. - The method is not affected by the relationships of the study variables.
Disadvantages	- Lower accuracy rates - Instability of the models derived - The decisions cannot be revised or improved (no backtracking technique). - A statistical significance of the variables is not provided	- It requires data pretreatment - Higher complexity for calculating importance rates - More time required for determining the relative importance of the variables (almost an hour) - Tedious procedure for determining the relative importance of the variables (additional methods must be applied). - A statistical significance of the variables is not provided