



Citizens' opinion about investment in public transport projects in cities

Francisco Calvo-Poyo, Adriana Medialdea & Ramón Ferri-García

To cite this article: Francisco Calvo-Poyo, Adriana Medialdea & Ramón Ferri-García (2020) Citizens' opinion about investment in public transport projects in cities, International Journal of Sustainable Transportation, 14:10, 806-818, DOI: [10.1080/15568318.2019.1630529](https://doi.org/10.1080/15568318.2019.1630529)

To link to this article: <https://doi.org/10.1080/15568318.2019.1630529>



Published online: 09 Jul 2019.



Submit your article to this journal [↗](#)



Article views: 101



View related articles [↗](#)



View Crossmark data [↗](#)



Citizens' opinion about investment in public transport projects in cities

Francisco Calvo-Poyo^a, Adriana Medialdea^b, and Ramón Ferri-García^c

^aDepartment of Civil Engineering, University of Granada, Granada, Spain; ^bE.T.S.I. Caminos, University of Granada, Granada, Spain;

^cDepartment of Statistics and Operations Research, University of Granada, Granada, Spain

ABSTRACT

In recent times, the increasing demand for mobility has given rise to new projects for improved public transport (PT) infrastructures. Such is the case of Andalucía (Spain) where, since 2004, the transit networks were improved and new underground and light rail systems were built. These investments were made without taking into account the opinion of the citizens. In this context, this paper expounds the results of a telephone survey conducted in 2015 to record citizens' opinions about the investment funneled into local transport infrastructures. The responses were analyzed through decision trees, in order to identify groups of homogeneous characteristics and common opinion. The results indicate that most citizens believed investment to have been insufficient, even though nearly half the respondents affirmed the PT infrastructures in their city were adequate. The results of the two majority opinion groups showed that: when citizens believe the PT infrastructures of their city are not adequate, and they are not satisfied with key attributes of the PT, they hold investment to be insufficient; in contrast, if citizens describe the PT infrastructure as adequate, they appraise key aspects and support the use of public funding for financing the PT, then they are satisfied with the investment made. Furthermore, some decision rules, and the most influential variables in the study problem (PT infrastructures adequacy, coverage, being a suburban bus user or not, and the support to use public funding for PT) were identified.

ARTICLE HISTORY

Received 28 August 2018

Revised 6 June 2019

Accepted 7 June 2019

KEYWORDS

urban transport; public transport investment; public transport survey; decision trees

1. Introduction

In recent years, many large-scale public transport (PT) projects have been carried out in urban areas to facilitate user mobility. Since 2004 (Calvo, Ferri, & Fernández, 2016), a number of vast PT projects have been undertaken in cities of Andalusia (southern Spain): improvements in the suburban rail networks of Seville and Cádiz, an underground in Seville, Light Rail Transit (LRT) systems for Seville, Vélez-Málaga, and Jaén, an improved bus system in Granada (similar to a Bus Rapid Transit), the Málaga underground, and Granada's LRT (the latter two under construction at the time of this study). The great magnitude of the economic resources needed for these transport projects can slow down their implementation for years, during which user demand may grow substantially. Therefore, citizens overall tend to find PT infrastructures to be inadequate, and investment insufficient. Worse yet, some of the aforementioned projects were failures (e.g., the LRT of Vélez-Málaga and Jaén are closed), or have generated strong social dispute (e.g., the new bus system and the LRT in Granada). Criticism is heated because all these projects were carried out without the participation of citizens. For this reason, in an aim to improve transport planning in future projects, it was deemed valuable to know the opinion of citizens regarding investment in local PT projects during the past 10 years.

In order to learn the perceptions of citizens about these PT projects, a telephone survey campaign was carried out in

the Andalusian "capital cities": Almería, Cádiz, Córdoba, Granada, Huelva, Jaén, Málaga, and Seville. The ultimate goal of the study was to determine the characteristics (socio-economic, habits of transport and opinion regarding PT) of the different opinion groups regarding investment. The responses were analyzed by means of decision trees, which allowed them to be classified according to specific characteristics of the participants. In addition, this methodology serves to highlight the conditions (concatenated characteristics) that increase the probability of belonging to one opinion group or another, and to identify the most influential variables of the study problem. The C5.0 algorithm was applied to later compare the level of accuracy of each tree through cross-validation, by setting different parameters of minimal cases in the terminal nodes. Use of this algorithm made it possible to create a scheme of the relationships between citizens' opinions about investment in PT and the conditioning factors that might influence such opinions. The procedure also allowed for decision rules to be extracted, so as to estimate the relevance of the exogenous variables considered in the study on the objective variable.

As main contributions of this study, it can be said that it is based on a survey that took in a great number of citizens (1,205 valid responses), analyzed using an advanced statistical method (decision trees). Above all, it provides knowledge regarding the opinion of citizens about the investment in PT in their cities. This is a matter that has largely been

overlooked (at least as the main objective) in previous studies.

The study is structured as follows. In the first place, the state of the art is reviewed. Next, the aim of the study is presented. Thirdly, the survey is presented, and the methodology is justified and described. Then, the data processing is explained. Afterward, the implementation methodology is applied to the survey responses, and the results are analyzed. Finally, the conclusions are expounded.

2. State-of-the-art

In recent decades, surveys have frequently been used to study citizens' perceptions about city transport projects. For example, Fan and Guthrie (2009) analyzed the local social effects of PT improvements by asking citizens how transport had improved their neighborhoods. In order to explain the factors influencing that opinion, the questionnaire included questions about their transport habits (mode choice, frequency of PT use, etc.), and their perceptions (quality) and attitude (reasons to use/not use PT, PT support, etc.) toward transit. Moreover, the questionnaire collected socioeconomic characteristics of the population (including age, gender, salary, education, and occupation). In their study, 60% responded that the neighborhood had improved after implantation of the new means of transport. The results proved useful to segment the types of zones where perceptions were more positive, and to study the characteristics of those citizens who best perceived transport improvements. Lawless (1999) examined the relationship between transport investment and urban regeneration in Sheffield (United Kingdom), where a substantial program of new LRT and improved roads had been carried out; hence, "before-and-after" surveys were conducted (interviews with Sheffield-based developers/city planners, local and national agents, questionnaires for visitors and residents, and telephone surveys of companies). The surveys addressed the positive impact of enhancing roads in terms of five elements of urban regeneration: image, property values, land use, business (re)location, and labor mobility. It was concluded that investment had a relatively limited impact on regeneration and, where effects could be identified, greater investment in roadways proved significantly more beneficial than did the LRT. Brown, Werner, and Kim (2003) undertook a survey in Salt Lake City to study the factors supporting the switch to transit use in the context of some changes in the transport system (parking shortages and the opening of a LRT). The survey included questions to gather citizens' perceptions regarding transit use and satisfaction, the effect of the new means of transport (livability, positive impacts), and their preference for the bus or the LRT. While 60% of the respondents thought the new LRT line improved city livability, 67% moreover thought it had a positive impact on the areas. The LRT was found to be much preferred over the bus.

According to Mostafavi, Abraham, and Vives (2014), assessment of public perceptions is a major component of gaining public support for developing sustainable

infrastructure financing policies. Their study, likewise based on a survey, looked into the determinant factors of public perceptions about innovative financing. The key questions governing financing were explained through economic indicators and aspects of the local infrastructure (including income per capita, population, unemployment rate, drinking water needs, roads in poor condition, and PT expenditure) as well as personal characteristics of the participants (gender, marital status, age, number of children, household annual income, number of cars in the household, distance to an interstate highway, and education).

Agrawal (2015) reviewed a set of 56 surveys regarding PT in the US, in order to identify similar items usually covered in these queries, and to detect general patterns in public opinion about transit that might emerge. Four topics were frequently found across the different surveys: the reasons why people support public transit; opinions about transit service quality; the extent to which people support improving transit as a general concept; and support levels for raising additional revenues to improve transit. Analysis showed that clear majorities believed transit to bring specific benefits to their community (congestion relief, positive impact on local economy, environmental issues, mobility options, etc.). There was also a clear majority support behind improvement of transit as a general concept. However, fewer people (one-third to one-half) supported the general concept of increased spending on transit, and considerably fewer than half supported raising taxes to increase transit funding.

Authors Hensher and Daniels (2011) show the results of a quarterly telephone survey of 1,000 Australians to monitor changes in the community's transport opinions, transport confidence, and reactions over time. Some key indicators included in this survey were: transport as a national priority, the highest priority issue within transport, government responsibility for transport, and private sector involvement in PT. In light of the national priorities considered (health, economy, employment, education, environment, law and order, infrastructures, housing, social services, security and transport), infrastructures were rated as having medium importance, while transport itself took the last place. Within the transport system priorities, 52% stated that PT should be improved as opposed to roads, railways or airports. Similarly, Luke and Heyns (2013) report on an annual survey of 1,000 South Africans to gauge opinion on transport-related matters. Their study compares the current PT policies in South Africa and public opinion on PT to see how well they are aligned. The questionnaire gathered demographic information such as location, age, gender and employment along with the key indicators related to transportation. Transport took third place as a national priority (very close to the second position). In addition, within the transportation system, the largest proportion of respondents (24.7%) assigned the greatest importance to improvement of the TP, over all the other options.

In a related study, Calvo-Poyo, Freiria, Medialdea, and Antunes (2018) gathered public opinion about the implantation of Bus Rapid Transit in Coimbra (Portugal) instead of the LRT that was already under construction, but it was

canceled after investing more than 100 million euros. The key questions surrounded the need for this new mode of transport and support for public spending on it. Other variables that were found to be significant in explaining response overall were: accepting the duration of construction, positive effects on the local economy, self-perception as a potential user, adequacy of the route, and gender. The main hypothesis of study—if citizens considered the new PT mode necessary, then supported the investment—was confirmed.

In short, surveys are commonly used to poll the opinion of citizens regarding PT projects in cities. In addition to questions about the socio-economic characteristics of citizens, the surveys usually include items about how these projects improve life in the neighborhoods, their positive impact, transportation habits, perceptions of transportation, and attitudes towards PT. However, few of the studies consulted contain as key questions the citizen's support for PT projects and the investment required, items that could reflect the role of citizens as participants in PT projects. Such participation might serve to improve transport planning, and fortify citizens' support of transport projects.

3. Study objective

The main objective of this study was to identify the characteristics—socio-economic characteristics, PT use habits, and opinions about various aspects of the PT system—of the population sectors that considered investment to be sufficient, or otherwise that it should have been greater. By doing this, the profile of the citizens believing there is a lack of investment emerges, along with the profile of those satisfied with the level of investment. Both aspects can be of great help in the planning of future PT projects in cities. Further objectives were to identify the aspects that most influenced public opinion regarding PT investment, and to discover chains of conditioning factors that could increase the possibility of citizens' satisfaction or dissatisfaction with investment.

4. The survey

The sample size of the survey was 1,205 (valid responses), which gave a sampling error of 2.88%. The survey was conducted in 2015 by telephone, using the computer-assisted interview system CATI. The rate of response was 45.9%, close to the mean response rate for interviews using this system. The survey included questions about socio-demographics of the participants, their habits of PT use, and their opinion regarding a variety of PT issues, including investment (Tables 1, 2, and 3).

One controversial issue in Andalusia in the past two decades is the choice between implementing an improved bus system (for example, with exclusive bus lanes) or a LRT (Calvo et al., 2016). This is reflected in the fact that most respondents (59.3%) believed that bus lanes had a positive impact in the city, whereas the LRT is largely considered not to have a positive impact (45.4%). Furthermore, LRT is

held to be the least adequate means of transport by 47.4% of respondents, while the bus is considered the best (82.1%). Finally, when queried about the level of investment in PT over the last 10 years, 72% responded that it was insufficient, while 28% expressed satisfaction with the investment in their city. A more detailed study of the results of the survey can be found in Calvo et al. (2016), whose main findings were that:

- People living in cities which already have an underground show a greater preference for underground means of transport than those who live in cities that have only surface transit.
- Most people believe that investment in PT infrastructures should have been greater.
- Urban and suburban buses are considered the most adequate transport modes for the Andalusian capital cities and the transport needs of their inhabitants.
- LRT is considered the least adequate transport mode.

5. Methodology

Given the nature of the problem at hand—discerning the common factors in groups of population having similar opinions regarding investment in PT through a highly dimensional data set—the use of classic statistical methods (e.g.) was discarded, and a machine learning model was adopted. It prioritizes the discernment of classes and the visualization of highly dimensional data through the extraction of the simplified structure. Moreover, among the available methods of machine learning, the best option was held to be decision trees, as a highly legible model affording high predictive precision.

Moreover, the differences among users in terms of their perceptions should be taken into account. Previous studies of consumer preferences present nonobservable heterogeneity due to the qualitative nature of certain items (PT support, adequacy of the mode of transport for the city, etc.), the different attitudes that passengers have towards the use of PT (i.e., reasons for not using PT), the different ways of viewing aspects of the service (PT perception, relative importance of the attributes of PT), their preferences (adequacy of the mode of transport for personal needs), and the social and economic characteristics of citizens (Cirillo, Eboli, & Mazzulla, 2011). This means that discrete choice models (e.g., logit) are not appropriate, since they assume homogeneity among individual responses.

Classification and regression trees, a nonparametric model family with no pre-defined underlying relationship between the target (dependent) variable and the predictors (independent variables), have been widely employed in business administration, agriculture, industry, and engineering. Enabling automatic searches for the best predictors and the best threshold values for all predictors to classify the target variable, CART has proven to be a powerful tool, particularly when dealing with prediction and classification problems (De Oña, De Oña, & Calvo, 2012) and therefore stood out as the most useful procedure for the present study.

Table 1. Socio-demographic variables.

Variable	Categories/range	% ($\pm\sigma$)	% ($\pm\sigma$) without NR/DK
<i>Socio-demographic variables</i>			
Gender	Male	46	46
	Female	54	54
Age ^a	[18, 79]	46.74 (± 15.84)	46.74 (± 15.84)
Hometown	Almería	8.6	8.6
	Cádiz	7.4	7.4
	Córdoba	12	12
	Granada	12.2	12.2
	Huelva	7.7	7.7
	Jaén	7.7	7.7
	Málaga	17.9	17.9
	Seville	26.5	26.5
	Unemployed	23.4	24.9
Salary	<1000 €	23.8	25.3
	1000–1500 €	20.2	21.5
	1500–2100 €	15.9	16.9
	>1500 €	10.7	11.4
	NR/DK	6	
Education	Primary education/no studies	8.1	8.1
	Secondary education	13.1	13.1
	Technical education	10.5	10.5
	Higher certificate	26.1	26.1
	University graduate	28	28
	University postgraduate or higher	14.2	14.2

Notes. PT: public transport. NR/DK: no response/do not know.

^aVariables included in the tree.

Table 2. Rail transit offer and PT demand variables.

Variable	Categories/range	% ($\pm\sigma$)	% ($\pm\sigma$) without NR/DK
<i>Rail availability</i>			
Underground	Yes	44.4	44.4
	No	55.6	55.6
LRT	No	53.6	53.6
	Yes	26.5	26.5
	Under construction	12.2	12.2
	Built but not operating	7.7	7.7
Suburban rail	Yes	51.8	51.8
	No	48.2	48.2
<i>PT use</i>			
Number of trips (working days)	[0,14]	1.25 \pm 1.9	1.25 \pm 1.9
Urban bus	No	55.2	55.2
	Yes	44.8	44.8
Suburban bus ^a	No	94.6	94.6
	Yes	5.4	5.4
Underground	No	95.5	95.5
	Yes	4.5	4.5
LRT	No	98.1	98.1
	Yes	1.9	1.9
Suburban rail	No	95.9	95.9
	Yes	4.1	4.1

Notes. PT: public transport. NR/DK: no response/do not know.

^aVariables included in the tree.

Decision trees furthermore constitute a classification technique (prediction of categorical data) that is easy to interpret while producing a high yield in most cases. The underlying notion is that of a tree with its respective branches (conditions) and leaves (nodes), which may be internal or terminal. The leaves or terminal nodes indicate the classes that should be assigned to the variables of study. Each internal leaf contains a variable that undergoes an evaluation, which will give rise to two conditions that divide the data into two sub-sets, one for each branch that comes out of that leaf.

Decision trees have numerous applications, but transportation engineering is one field where their use has become

prolific in recent years. They were applied by Kashani and Mohaymany (2011) to identify the most influential risk factors in the severity of injuries in traffic accidents, using the CART type of decision trees, with regression and classification. This particular method lends the possibility of obtaining a measure of the importance of each variable based on the amount of information it contributes when added to the decision tree. In the area of perception of quality of service, this technique was applied in the studies by De Oña et al. (2012), De Oña, Eboli, and Mazzulla, (2014), and Hernández, Monzón, and De Oña (2016). Moreover, in De Oña et al. (2012, 2014), the resulting tree was used to elaborate decision rules that made it possible to associate an *outcome* with a series of values given by certain

Table 3. PT support and perception variables.

Variable	Categories/range	% ($\pm\sigma$)	% ($\pm\sigma$) without NR/DK
<i>PT funding and support</i>			
Knows that PT price is subsidized	Yes	39.7	39.7
	No	60.3	60.3
PT should be promoted over private transport	Yes	97.3	97.3
	No	2.7	2.7
Use of public funding for PT support*	Little support	3.9	3.9
	Neutral opinion	17.5	17.5
	Great support	78.6	78.6
	NR/DK	10.2	
Investment in PT*	Insufficient	64.6	72
	Sufficient	25.2	28
	NR/DK	10.2	
<i>Vehicles should be replaced due to their age</i>			
Urban bus	No	34.2	56.1
	Yes	26.7	43.9
	NR/DK	39.1	
Suburban bus	No	55.2	90.6
	Yes	5.7	9.4
	NR/DK	39.1	
Underground	No	60.8	99.9
	Yes	0.1	0.1
	NR/DK	39.1	
LRT	No	60.7	99.7
	Yes	0.2	0.3
	NR/DK	39.1	
Suburban rail	No	57.8	95
	Yes	3.1	5
	NR/DK	39.1	
<i>Reasons for not using PT</i>			
Coverage	No	80	80
	Yes	20	20
Travelling or waiting time	No	56.8	56.8
	Yes	43.2	43.2
Lack of comfort	No	77	77
	Yes	23	23
Ticket prices	No	77.6	77.6
	Yes	22.4	22.4
Unsafe	No	96.1	96.1
	Yes	3.9	3.9
Other	No	73.4	73.4
	Yes	26.6	26.6
NR/DK	No	89.4	89.4
	Yes	10.6	10.6
<i>Is the means of transport oversized?</i>			
Urban bus	No	80.7	80.7
	Yes	19.3	19.3
Suburban bus	No	96.8	96.8
	Yes	3.2	3.2
Underground	No	95.8	95.8
	Yes	4.2	4.2
LRT	No	97.8	97.8
	Yes	2.2	2.2
Suburban rail	No	97	97
	Yes	3	3
NR/DK	No	91.1	91.1
	Yes	8.9	8.9
<i>PT perception</i>			
PT helps improve city life	No	3.6	3.6
	Neutral opinion	6.4	6.4
	Yes	90	90
PT infrastructures are adequate ^a	No	26.4	26.4
	Neutral opinion	31	31
	Yes	42.5	42.6
	NR/DK	0.1	
Frequency	Poor	24.6	24.8
	Acceptable	31	31.3
	Good	43.5	43.9
	NR/DK	0.9	
Comfort	Poor	17.6	18
	Acceptable	28	28.6
	Good	52.2	53.4
	NR/DK	2.2	
Safety *	Poor	6.5	6.5
	Acceptable	14.2	14.3
	Good	78.7	79.2

(continued)

Table 3. Continued.

Variable	Categories/range	% ($\pm\sigma$)	% ($\pm\sigma$) without NR/DK
Coverage ^a	NR/DK	0.6	
	Poor	27.8	28.6
	Acceptable	20.1	20.6
	Good	49.5	50.8
PT quality/price ratio ^a	NR/DK	2.6	
	Poor	46	47.7
	Reasonable	49.3	51
	Good	1.2	1.3
Positive impact of bus lane projects	NR/DK	3.5	
	No	16.7	18.2
	Neutral opinion	20.6	22.5
	Yes	54.4	59.3
Positive impact of LRT projects	NR/DK	8.3	
	No	22.4	45.4
	Neutral opinion	9.1	18.5
	Yes	17.8	36.1
<i>Adequacy of mode of transport for the city and personal transport needs</i>			
Urban bus	NR/DK	50.7	
	Poor	5.5	5.5
	Neutral opinion	12.4	12.4
Suburban bus	Good	82.1	82.1
	Poor	10.2	10.2
	Neutral opinion	16.4	16.4
Underground	Good	73.4	73.4
	Poor	42.8	42.8
	Neutral opinion	14	14
Light rail	Good	43.2	43.2
	Poor	47.4	47.4
	Neutral opinion	15.4	15.4
Suburban rail	Good	37.2	37.2
	Poor	17.1	17.1
	Neutral opinion	18.1	18.1
<i>Importance of attributes for using PT</i>			
Comfort	Good	64.8	64.8
	Little importance	5.7	5.7
	Neutral opinion	19.5	19.5
Age of vehicles ^a	Great importance	74.8	74.8
	Small	12.5	12.5
	Neutral opinion	20.2	20.2
Safety	Great	67.3	67.3
	Small	6.1	6.1
	Neutral opinion	14.2	14.2
Frequency	Great	79.7	79.7
	Small	7.5	7.5
	Neutral opinion	11.2	11.2
Travel time	Great	81.3	81.3
	Small	6.6	6.6
	Neutral opinion	17.7	17.7
Ticket prices	Great	75.7	75.7
	Small	10.5	10.5
	Neutral opinion	22.5	22.5
Coverage	Great	67	67
	Small	7	7
	Neutral opinion	16.7	16.7
	Great	76.3	76.3

Notes. PT: public transport. NR/DK: no response/do not know.

^aVariables included in the tree.

variables or *incomes*, similar to the way results are obtained using association rules. Another type of decision tree is the one derived from the ID3 algorithm (Quinlan, 1979), and more recently the algorithms C4.5 and C5.0 (Quinlan, 1993), involving different metrics to build the trees. Authors Wong and Chung (2007) applied the C5.0 algorithm to model satisfaction with the quality of service among airline passengers in Taiwan.

An additional field of application for decision trees in transportation studies is the modal choice. Arentze and Timmermans (2004) used a CHAID decision tree induction method to derive decision trees from activity-travel choice

behavior. Along similar lines, Janssens et al. (2006) showed that integrated Bayesian networks and decision trees could be used for modeling activity-based transportation with better predictive power than CHAID decision trees. Sekhar, Minal, and Madhu (2016) modeled the mode choice behavior of commuters in Delhi by means of Random Forrest, which takes the average of a collection of decision trees. Xie, Lu, and Parkany (2003) investigated the capability and performance of travel modal choices modeling using decision trees, specifically the C4.5 learning algorithm, and a neural network.

C5.0 trees were chosen for the present study because of their advantages regarding precision, computation speed,

and simplicity of the resulting trees when compared to other methods. The choice of the variable to be split is made in view of the criteria for maximum extraction of information from the existing data. To this end, the C5.0 algorithm uses a criterion of information gain based on entropy coefficients, in which p_i denotes the probability of class i , among m possible classes of the objective variable, with entropy defined as:

$$H(p) = - \sum_{i=1}^m p_i \log p_i$$

Thus, maximum entropy is reached when the classes are equally probable, and the uncertainty about the choice of the correct class is maximized. The information gain is defined as the difference in entropy before and after the partition of the classes given certain fixed values of an attribute. Let S , with k partitions, be a division of the classes of a specific variable, and let n_i be the number of classes of the i -eth partition, so that $n = \sum_{i=1}^k n_i$; then, the gain in information has the following expression:

$$IG(S) = - \sum_{i=1}^m p_i \log p_i - \sum_{i=1}^k H(S_i) \frac{n_i}{n} = H(S) - \sum_{i=1}^k H(S_i) \frac{n_i}{n}$$

The variable to be divided at each one of the levels of the tree is therefore the one maximizing the information gain. In other words, it is the one having a high entropy before division and a low entropy or uncertainty after the division. The process continues for smaller subsets, and once this process is completed, the branches that do not contribute sufficiently to the purity of the classification are pruned. Pruning ensures that overfitting of the model is avoided.

In this final stage, each subtree is explored individually and the error of classification is estimated with and without each one. Let e_{-s} be the rate of error eliminating subtree s , and let e be the rate of error of the complete tree; then, the following limit is set to establish the formula:

$$b_e = e + z_{\alpha/2} \sqrt{\frac{e(1-e)}{n}},$$

so that if $b_{e_{-s}} > b_e$, the subtree is pruned.

Decision trees provide a series of goodness-of-fit measures. The one most used is *accuracy*, but it is also important to bear in mind the marginal precision of the algorithm—that is, its capacity to properly classify each individual of the different classes that exist. In the case of binary classification, a distinction is made between *sensitivity* and *specificity*, measuring the capacity of prediction of the minor and major classes, respectively. The formulas for calculating these are:

$$\text{Precision} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

where TP (true positive) is the number of cases correctly classified in the major class, FP (false positive) is the number of cases incorrectly classified in the major class, TN (true negative) is the number of cases correctly classified in the minor class and FN (false negative) is the number of cases incorrectly classified in the minor class. Another tool for measuring goodness-of-fit is the confusion matrix, wherein each row denotes the actual observation of each class, while each column represents the predicted level. However, one of the measures that best reflects the performance of the model is AUC (Area Under the ROC curve), because it is able to reflect in a single magnitude the rates of true and false positives (TP and FP, respectively), defined as:

$$AUC = \frac{1 + TP - FP}{2}$$

A perfect model would have an AUC equal to 1, meaning that the model allows for a total separability of the classes. In turn, poor models would have values near 0.

The measurements obtained are calculated over the data used to construct the model. Hence these measures tend to be overly optimistic regarding the goodness-of-fit that the model will have when it is applied to a new data set. To resolve this discrepancy, the literature has proposed *k-fold cross-validation* (CV). This method consists of dividing the set of data upon which the model is going to be built or trained into k parts. Then, the model is constructed on $k - 1$ partitions, and evaluation is made of the partition that was not used in the construction. The process is repeated for all the partitions until k measures of goodness-of-fit are obtained; they provide an approximation that is closer to reality, from which the arithmetic means can be derived as the final approximation.

The number of partitions chosen will give rise to one or another type of cross-validation, and the ones most widely used are 5-fold CV, 10-fold CV, and leave-one-out CV. The 10-fold cross-validation has been shown to be optimal for classification of algorithms despite the fact that, in practice, it is a biased estimator given its instability (as the partitions are made at random, they depend on a random seed). Leave-one-out cross-validation divides the set of data into n partitions (n being the size of the database), for which reason with each step the model is built over $n - 1$ data and it is evaluated over the remaining instance. The latter method, while entailing greater computational costs, is scarcely affected by bias since there is no random element, hence the estimation of error would be closer to the real error when compared with the other options for cross-validation (Kohavi, 1995; Cawley & Talbot, 2003). The smallest variance would therefore correspond to leave-one-out cross validation (Kearns & Ron, 1999).

6. Data analysis

6.1. Preprocessing

First, the data were cleaned, eliminating the instances presenting abnormal behavior, and the instances and variables presenting an excessive percentage of missing values

Table 4. Confusion matrix.

		Reference	
		Sufficient	Insufficient
Prediction	Sufficient	161	18
	Insufficient	72	56

(>50%). Individuals who responded “Do not know/no answer” were also removed from the sample. Then the noisy data were removed by means of machine learning techniques, namely a robust ensemble filter based on the classifier C4.5 (Verbaeten, 2002). This preprocess gave 983 respondents as the final number of cases.

6.2. Building the decision tree

Construction of the model entailed training the complete data set using the C5.0 algorithm and 10-fold cross validation as the resampling method, establishing a minimum of 14 cases for each inner node and a confidence level for pruning of $\alpha = 0.25$. Cross-validation with 10 iterations for resampling and adaptative boosting with 20 trials were applied as well. The latter procedure helps improve the accuracy of the tree and consists of generating as many classifiers as trials, so that each one of them assigns a vote to the predicted class, and at the end of the process the votes given by each classifier are counted to determine the correct class (Freund & Schapire, 1996).

6.3. Validation

The result of the training was compared with the result given by the same model over a subset of the whole data previously reserved for this testing phase (30% of the whole sample, as the remaining 70% was used in the training). The AUC obtained was 87.4%, with an accuracy of 70.6%, sensitivity of 69.2%, and specificity of 75.6%, respectively.

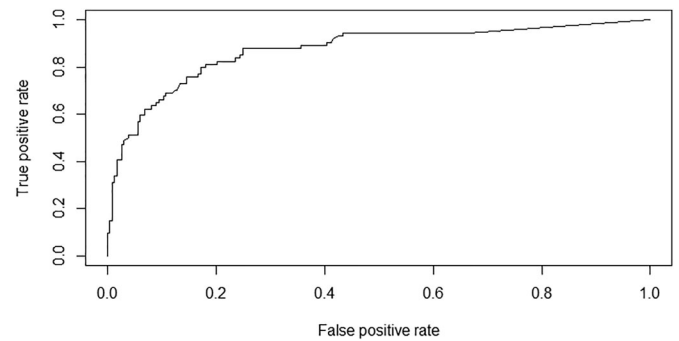
The confusion matrix (Table 4) shows the certainties and the errors of prediction of the model on the test set, where the rows represent the number of predictions for each class, and the columns are the real classes; in this way, the number of correct classifications predicted by the model is found in the main diagonal of the matrix.

The ROC curve is the graphic equivalent of the AUC validation measure, interpreted as the area found below the curve (Figure 1). It represents the rate of true positives with respect to false positives, so that a greater area under the curve means a better performance of the model. According to this, the ROC curve (Figure 1) shows the great performance of the model.

7. Results and discussion

7.1. Decision tree

Once the data were processed according to the steps described in section 6, eight variables were finally selected for the tree (Figure 2): adequacy of infrastructures, coverage, use of suburban buses, support for PT, security, age,

**Figure 1.** ROC curve.

importance attributed to the age of the vehicles, and value for money. As can be seen, these variables are mainly related to the perception of PT.

Starting with the respondents who mostly consider (more than 50% in each node) that investment was insufficient, and continuing this analysis depending on the population of each of the nodes, the following comments can be made.

- Node 2 contains the largest group of respondents (550 respondents, 56.1% of the sample); of these, 89.5% considered that investment in local PT was insufficient. This group of citizens either finds PT infrastructures in their city to be inadequate, or expresses a neutral opinion about them. Yet some discerning characteristics of the respondents included in Node 2 deserve mention here. As to their reasons for not using PT, 54.4% argued too much travel time and waiting (43.2% in the overall profile). Meanwhile, 76.5% affirm that PT frequency is poor (24.6% of the whole data set) and 43.8% of them are dissatisfied about coverage (28.6% in the overall profile). In addition, there is a greater perception of a poor quality/price ratio for PT (55.3% qualify it as poor, as compared with 47.8% in the overall profile of respondents). To sum up, people in this node are highly dissatisfied with key PT quality attributes, and may be expressing a demand for investment and improvements in travel time, frequency and coverage.
- The second largest node of respondents that held investment in local PT to be insufficient (67.2% of them) is Node 15, with 74 respondents (7.5%). This node includes people under the age of 73 and non-users of suburban buses. They consider the TP infrastructure, safety and coverage adequate (or have a neutral opinion regarding the latter aspect), and support investment in PT. At the same time, they lend importance to the age of the PT vehicles, and disapprove of the PT quality/price ratio. Yet the members of this node use PT more frequently (1.5 trips per day, compared to 1.25 on average). Moreover, they positively appraise PT frequency (59.7% vs. 43.2% in the general case), comfort (65.7% vs. 53.4%), and the impact of bus lanes (78% vs. 59.3%) as well as LRT (58.3% vs. 36.1%). Therefore, in this node we find frequent users of PT, who have a high opinion about it and its positive impact, but who also believe that more expenditure is necessary to improve local PT.

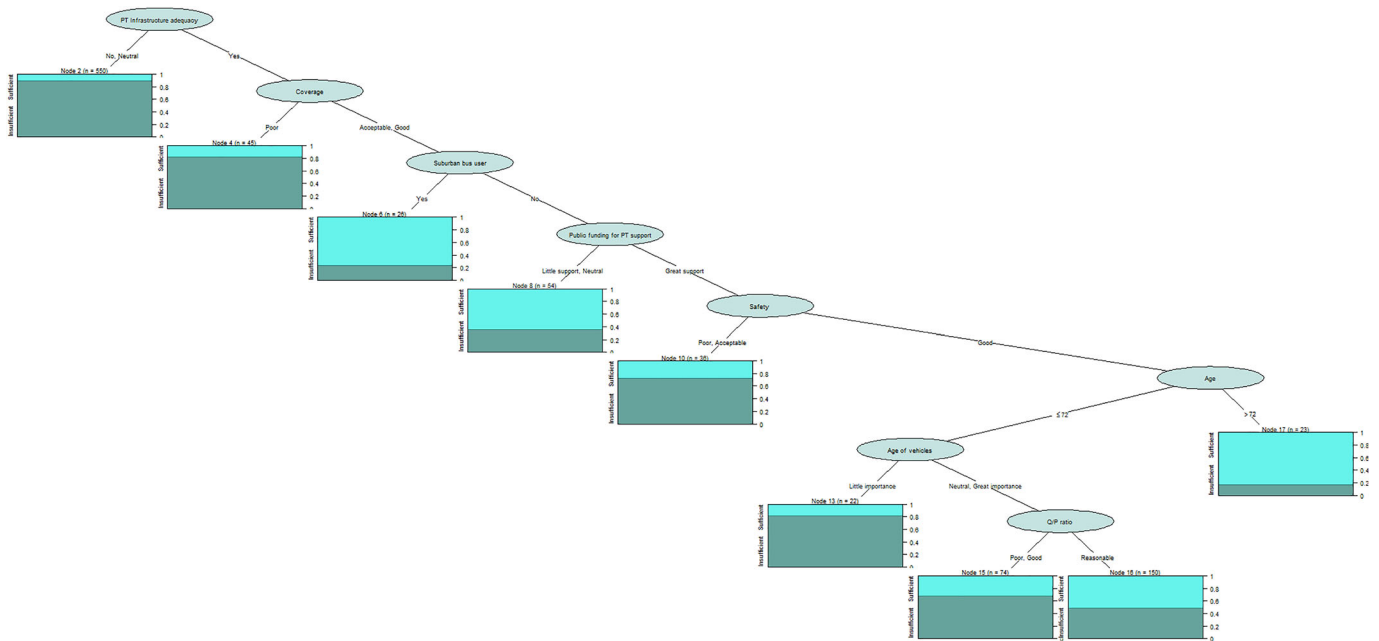


Figure 2. Decision tree for investment in PT.

- Node 4 (which includes 45 respondents, 4.6% of the sample) indicates that when citizens have a good opinion about PT infrastructures but are dissatisfied with coverage, they tend to conclude that the investment in PT was insufficient (82.2% of them). The differences between this node and whole sample are remarkable in that there is a higher proportion (38.8% compared to 25.3% on the average) of salaries below 1,000 euros. Besides, 11.1% of these particular respondents do not believe that PT improves life in cities (3.6% in the overall profile).
- In Node 10 there are non-users of the suburban bus (36 respondents, 3.7% of the sample) who consider infrastructures adequate, do not find coverage bad, consider it important to invest in PT, and believe PT is unsafe (or have a neutral opinion). In this node, 72.2% of respondents consider investment insufficient. This node has a greater female presence (66.7%, as compared to 54%) and low salaries predominate (47.1% vs. 25.3%). Although people in this group appraise the frequency and comfort of PT as adequate (62.2% of agreement in both cases versus 43.9% and 53.4%, respectively), only 77.8% believe it improves city life (90% in the general case).
- Finally, Node 13 groups respondents under 73 years of age who hold infrastructures and PT safety to be adequate, have a positive or neutral opinion on coverage, do not use the suburban bus, consider investment in PT important, and do not consider the age of the vehicles to be important. People (81.8%) in this node respond that investment is insufficient. The predominant profile is unemployed (42.1%, versus 24.8% on average), regular PT user (68.2% vs. 52.9%) and high satisfaction with frequency (77.3% vs. 43.2%) and comfort (72.7% vs. 53.4%).
- The largest group (Node 16) (150 respondents, equal to 15.2% of the total) who mostly consider that investment in PT was sufficient (50.7% satisfied with the investment) is made up of non-users of suburban buses under the age of 73. These citizens hold the infrastructures and security of PT to be adequate, have a positive or neutral opinion about coverage, support investment in PT, consider the age of the vehicles as important, and the quality/price ratio acceptable. In this node, frequency and comfort obtain more positive evaluations than in the global sample: respectively, 76.3% as compared to 43.2%, and 86.8% compared to 53.4%.
- Node 8 contains those respondents who are not suburban bus users, consider infrastructures adequate, have a positive or neutral opinion about coverage, and show little support for investment in PT. In this case, 65.5% consider investment to have been satisfactory. Unemployment is greater in this node (30.4% vs. 24.8%), and PT use decreases (69.1% are nonusers, vs. 61.3%). Notwithstanding, frequency and comfort are more positively perceived than in the whole sample (59% vs. 43.2%, and 81.7% vs. 53.4%, respectively).
- Node 6 groups the suburban bus users that consider the infrastructures appropriate and have a positive or neutral opinion about coverage. Satisfaction with regard to the investment increases, reaching 76.9%. In this node unemployment predominates (35% vs. 24.8%) along with those having university studies (39.1% vs. 28%). A very frequent use of PT is observed, with an average of 2.7 trips per day. Moreover, the percentages of urban bus users and suburban rail users are much greater in this node (88.5% vs. 44.8%, and 34.6% vs. 4.1%, respectively). Finally, a greater satisfaction with some aspects of PT is observed, namely: improvement of life in cities (100% vs. 90%), frequency (73.9% vs. 43.2%) and safety (95.7% compared to 79.2%), and the projects of bus lanes (82.6% vs. 59.3%) and LRT (72.2% vs. 36.1%). Thus, it could be generally said that this node takes in people with low salaries, who are habitual users of suburban PT

Taking a closer look at the groups of respondents who mostly think that the investment was sufficient, and from a larger to smaller group size, it can be said that:

Table 5. Decision rules.

Rule				
If	Then	Accuracy (%)	Lift coefficient	% Respondents (% error ^a)
(1) Non suburban bus users, considering PT infrastructures adequate, coverage good or acceptable, and small or neutral support for PT investment	PT investment sufficient	67.2	2.6	3.9 (0.02)
(2) Respondents who consider the safety poor and show a great support for PT investment	PT investment insufficient	95.3	1.8	5.3 (0.48)
(3) Respondents who consider the PT infrastructure inadequate	PT investment insufficient	98.5	1.8	24 (1.37)
(4) Non users of the suburban bus who consider frequency, safety, and PT infrastructures adequate and lend importance to the age of the vehicles	PT investment sufficient	71.8	2	14.7 (0.34)
(5) Age above 29, considering the frequency inadequate, the urban bus adequate and having a non-neutral opinion (either positive or negative) about the suburban bus	PT investment insufficient	97.9	1.8	12.7 (1.03)
(6) Age above 43, considering the frequency and the urban and suburban bus modes adequate, not giving little importance to the age of vehicles and the PT infrastructures adequate	PT investment sufficient	63.9	1.7	9.5 (0.44)
(7) Respondents who consider PT infrastructures adequate and express minor or neutral support of the investment in PT	PT investment sufficient	60.4	1.8	4.4 (0.06)
(8) Citizens considering the price of PT reasonable, the frequency, PT infrastructure and urban bus mode adequate and having a good opinion about safety and coverage	PT investment sufficient	64%	2.1	8.4 (0.31)

(a) % Error: % Difference between whole data set and training set.

Table 6. Importance of the variables.

Variable	Importance (%)
PT infrastructures adequacy	99.69
Coverage	44.05
Suburban bus user	39.47
Use of public funding for PT support	36.83
Safety	31.23
Age	27.57
Age of vehicles	25.23
PT quality/price ratio	22.99

modes, with a positive opinion regarding PT services and investment.

- Lastly, Node 17 includes respondents over 72 who consider infrastructures and PT safety adequate, have a positive or neutral opinion on coverage, do not use the suburban bus, and support investment in PT. This node reflects the highest percentage of satisfaction with investment (82.6%). An increase in the use of PT (2.4 trips per day) and preference for the urban bus (65.2% vs. 44.8%) are also observed. Respondents in this node largely

consider the quality/price ratio acceptable (72.2% vs. 50.8%) and show a favorable opinion about frequency and comfort (78.3% vs. 43.2%, and 82.6% vs. 53.4%, respectively).

7.2. Decision rules

By building trees using the model C5.0, useful decision rules can be elaborated. They are not necessarily derived from the final branches towards terminal nodes of the optimal tree—they might also be sub-trees, or even different combinations than the ones obtained from the tree in question. These rules allow for a somewhat more in-depth interpretation of the trees.

The index of increase, or *Lift coefficient*, is a coefficient utilized as a metric in the rules of association (Brijs, Vanhoof, & Wets, 2003). It serves to measure how much an antecedent increases the probability of occurrence of a consequent, by means of the quotient between probability (accuracy) of the rule appearing in the dataset and the product of the percentages with which the antecedent and the consequent are

present in the set of data. For a rule of the type “if X , therefore Y ”, the *Lift coefficient* would be calculated as:

$$\text{Lift} = \frac{P(X \cap Y)}{P(X)P(Y)}$$

Therefore, it is a key coefficient when interpreting the relevance that a given rule may have for the set of data. A lift value equal to 1 indicates independence between attributes, while a lift value larger than 1 implies a certain degree of relation between them. In order to identify the sets of attributes having the strongest association with investment satisfaction, a threshold for the *Lift coefficient* was established. Accordingly, only the decision rules with a *Lift coefficient* over 1.5 were taken into account (Table 5).

Table 5 shows the most significant decision rules:

1. If citizens are not suburban bus users (96%), they consider infrastructures adequate, coverage good or acceptable, and do not firmly support the use of public resources for funding PT, there is an increased probability (160%) of being satisfied with the investment in PT.
2. If citizens support the investment in PT, but consider that safety is poor, the probability that they consider investment insufficient increases by 80%.
3. If citizens consider that the PT infrastructures of their city are inadequate, then the probability that they hold the investment to be insufficient increases by 80%.
4. If non-users of the suburban bus think that PT infrastructures, frequency, and safety are adequate, and consider the age of the vehicles as an important attribute when using PT, the possibility that they estimate investment in PT to be sufficient increases by 100%.
5. If citizens over 29 years of age rate the frequency as inadequate, urban buses as an adequate mode of transportation for their city, and do not have a neutral opinion about the adequacy of buses for suburban trips, the probability that they are dissatisfied with investment increases by 80%.
6. If citizens over 43 rate frequency as good and think that urban and suburban buses are a good option for their urban environment, and do not lend minor importance to the age of the vehicles when using PT, the chances of being satisfied with investment increase by 70%.
7. If citizens estimate that PT infrastructures are adequate and do not support investment in PT, the probability of considering investment as sufficient increases 80%.
8. If citizens appraise infrastructures, urban buses, price, frequency, safety and coverage as adequate, the probability that they are satisfied with the investment increases by 110%.

Below, Table 6 shows the importance of the variables included in the tree derived from the criterion of informational gain, using algorithm C5.0.

Table 6 shows the importance of the variables included in the tree to explain the dependent variable. Accordingly, the most relevant variables for explaining the opinion of citizens regarding investment in PT are their opinion on the adequacy of the transport infrastructures in their city,

coverage, being a suburban bus user or not, support for using public funding for PT, and safety.

8. Conclusions

As a major contribution of this study we would underline the fact that the survey covered a broad sector of the population (later analyzed by means of decision trees) who responded to the key question about the level of investment in PT in their city. The literature consulted does not have studies with this main objective and this realm of study. Deserving special mention are the findings listed below.

1. Nearly half the participants (47%) use PT on a regular basis, with a mean of 1.25 trips per person and work day. The urban bus is the mode of transport most utilized (for 45% of trips).
2. The vast majority of respondents believe that PT improves life in the city (90%), that PT should be fomented and encouraged over private transport (97%), and they support the use of public funding for PT (79%). The values obtained for these items are greater than similar indicators found in the literature review studied elsewhere and published previously. It can therefore be stated that the overall attitude toward PT expressed was very favorable.
3. Nearly half the respondents (43%) believe that the PT infrastructures in their city are adequate. Even so, most (72%) are of the opinion that investment has been insufficient. This proportion of citizens in favor of greater investment is also higher than the figures reported in the literature consulted. The decision trees made it possible to identify groups of population having homogeneous characteristics, but with a different opinion regarding the level of investment. The characteristics of the most numerous opinion groups are listed below:
 4. The largest group (56% of the sample) shows that when citizens consider the PT infrastructures of their city to be inadequate, they tend to consider investment as insufficient. These respondents are dissatisfied with key PT attributes—travel time, frequency, coverage and quality/price ratio—which might justify their demand for greater investment. Therefore, one practical recommendation would be to invest more in new vehicles and in extending the lines to improve public opinion of PT and, accordingly, approval about the level of investment. Moreover, better communication to the public about the benefits of city transport projects in view of the respective investment could be perceived as advantageous.
 5. The second most important group (15%) makes manifest that, if citizens express a good opinion about the PT infrastructures, coverage and safety, consider the quality/price ratio acceptable, support investment in PT, and lend importance to the age of the vehicles, then they tend to be satisfied with the investment made. Thus, proper PT infrastructure planning, along with

communication of its benefits to local inhabitants, and maintaining the quality of the transport services provided, can help citizens tend to believe that investment in PT is acceptable.

The decision rules reveal sequences of conditioning factors that have a decisive influence on public opinion regarding the level of investment in PT. The following are the most illustrative:

6. If citizens consider that the PT infrastructures of their city are inadequate, then the probability that they consider investment to be insufficient increases by 80%.
7. If citizens appraise infrastructures, urban buses, price, frequency, safety and coverage as adequate, the probability that they are satisfied with the investment increases by 110%.
8. If non-users of suburban buses (96% of the data set) consider infrastructures adequate, and coverage good or acceptable, the probability of considering the corresponding investment as satisfactory increases by 160%, despite not firmly supporting the use of public resources for funding PT.

Considering the relationship between the key questions included in a previous study (Calvo-Poyo et al., 2018) and the items on the survey used here, it can be affirmed that conclusions 5, 7, and 8 hold, and the importance of the variables (highlighted below) is in line with the hypothesis verified in that study.

Lastly, the aspects bearing a greater impact on citizens' opinion about the level of investment in PT in their city are the appraisal of the adequacy of the transport infrastructures in their city, coverage, being a suburban bus user or not, support for using public funding for PT, and safety. This underlines the importance of sound planning strategies when developing urban transport infrastructures and services, as well as clear communication to the public of the benefits of such projects.

Acknowledgments

This work was supported by the European Regional Development Fund (ERDF) of European Union, and the Public Works Agency and Regional Ministry of Public Works and Housing of the Regional Government of Andalusia (Spain) (AOPJA) under Grant G-GI3003/IDIE via the Researching Project "Tool for the preliminary evaluation of PT infrastructures".

Disclosure statement

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Agrawal, A. W. (2015). What Do Americans Think about Public Transit? A Review of U.S. Public Opinion Polling Survey Questions. Mineta Transportation Institute Publications. Retrieved from <https://transweb.sjsu.edu/sites/default/files/1132-US-public-opinion-on-transit-a-survey-review.pdf>

- Arentze, T., & Timmermans, H. (2004). A learning-based transportation oriented simulation system. *Transportation Research Part B: Methodological*, 38(7), 613–633. doi:10.1016/j.trb.2002.10.001
- Brijs, T., Vanhoof, K., & Wets, G. (2003). Defining interestingness for association rules. *International Journal "Information Theories and Applications"*, 10(4), 370–375.
- Brown, B. B., Werner, C. M., & Kim, N. (2003). Personal and contextual factors supporting the switch to transit use: Evaluating a natural transit intervention. *Analyses of Social Issues and Public Policy*, 3 (1), 139–160. doi:10.1111/j.1530-2415.2003.00019.x
- Calvo, F. J., Ferri, R., & Fernández, J. (2016). *Investment in PT projects in cities: What do citizens think about it?* Paper presented at CIT2016 – XII Conference on Transport Engineering, Valencia, Spain University of Valencia.
- Calvo-Poyo, F., Freiria, S., Medialdea, A., & Antunes, A. (2018). Citizens' survey for the implementation of a new means of transport. *Transportation Research Procedia*, 33, 251–258. doi:10.1016/j.trpro.2018.10.100
- Cawley, G., & Talbot, N. (2003). Efficient leave-one-out cross-validation of kernel Fisher discriminant classifiers. *Pattern Recognition*, 36(11), 2585–2592. doi:10.1016/S0031-3203(03)00136-5
- Cirillo, C., Eboli, L., & Mazzulla, G. (2011). On the asymmetric user perception of transit service quality. *International Journal of Sustainable Transportation*, 5(4), 216–232. doi:10.1080/15568318.2010.494231
- 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(12), 11164–11171. doi:10.1016/j.eswa.2012.03.037
- 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. doi:10.3846/16484142.2014.898216
- Eboli, L., & Mazzulla, G. (2011). A methodology for evaluating transit service quality based on subjective and objective measures from the passenger's point of view. *Transport Policy*, 18(1), 172–181. doi:10.1016/j.tranpol.2010.07.007
- Fan, Y., & Guthrie, A. (2009). *Assessing Neighborhood and Social Influences of Transit Corridors*. CTS Report 12–19 University of Minnesota. Retrieved from <http://www.cts.umn.edu/Publications/ResearchReports/reportdetail.html?id=2162>
- Freund, Y., & Schapire, R. E. (1996). *Experiments with a New Boosting Algorithm*. Paper presented at the Thirteenth International Conference on Machine Learning.
- Hensher, D., & Daniels, R. (2011). *Monitoring community views on transport confidence over time: The quarterly Transport Opinion Survey (TOPS)* (Vol. 34). Adelaide, South Australia: Australasian Transport Research Forum (ATRF).
- Hernández, S., Monzón, A., & De Oña, R. (2016). Urban transport interchanges: A methodology for evaluating perceived quality. *Transportation Research Part A*, 84, 31–43. doi:10.1016/j.tra.2015.08.008
- Janssens, D., Wets, G., Brijs, T., Vanhoof, K., Arentze, T., & Timmermans, H. (2006). Integrating Bayesian networks and decision trees in a sequential rule-based transportation model. *European Journal of Operational Research*, 175(1), 16–34. doi:10.1016/j.ejor.2005.03.022
- Kashani, A., & Mohaymany, A. (2011). Analysis of the traffic injury severity on two-lane, two-way rural roads based on classification tree models. *Security Science*, 49, 1314–1320. doi:10.1016/j.ssci.2011.04.019
- Kearns, M., & Ron, D. (1999). Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Computation*, 11(6), 1427–1453. doi:10.1162/089976699300016304
- Kohavi, R. (1995). *A study of cross-validation and bootstrap for accuracy estimation and model selection*. Paper presented at the International Joint Conference on Artificial Intelligence (IJCAI).
- Lawless, P. (1999). Transport investment and urban regeneration in a Provincial City: Sheffield, 1992–96. *Environment and Planning C: Government and Policy*, 17(2), 211–226. doi:10.1068/c170211
- Luke, R., & Heyns, G. (2013). PT policy and performance: The results of a South African public opinion poll. *Journal of Transport and Supply Chain Management*, 7(1), 8. doi:10.4102/jtscm.v7i1.96

- Mostafavi, A., Abraham, D., & Vives, A. (2014). Exploratory analysis of public perceptions of innovative financing for infrastructure systems in the United States. *Transportation Research Part A*, 70, 10–23. doi:[10.1016/j.tra.2014.10.002](https://doi.org/10.1016/j.tra.2014.10.002)
- Quinlan, J. R. (1979). *Iterative Dichotomizer 3 (ID3)* Stanford University. CA: Artificial Intelligence Laboratory.
- Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Mateo, CA: Morgan Kaufmann Publisher Inc. doi:[10.1007/BF00993309](https://doi.org/10.1007/BF00993309).
- Sekhar, R., Minal, C., & Madhu, E. (2016). Mode choice analysis using random forrest decision trees. *Transportation Research Procedia*, 17, 644–652. doi:[10.1016/j.trpro.2016.11.119](https://doi.org/10.1016/j.trpro.2016.11.119)
- Verbaeten, S. (2002). *Identifying mislabeled training examples in ILP classification problems*. Paper presented at the 12th Belgian-Dutch Conference on Machine Learning. Utrecht, The Netherlands, pp. 71–78.
- Wong, J.-Y., & Chung, P.-H. (2007). Managing valuable Taiwanese airline passengers using knowledge discovery in database techniques. *Journal of Air Transport Management*, 13(6), 362–370. doi:[10.1016/j.jairtraman.2007.07.001](https://doi.org/10.1016/j.jairtraman.2007.07.001)
- Xie, C., Lu, J., & Parkany, E. (2003). *Work travel mode choice modeling with data mining: Decision trees and neural networks*. Paper presented at the 82nd Transportation Research Board Annual Meeting, January 12–16, 2003. Washington, D.C.