UNIVERSIDAD DE GRANADA

E.T.S. INGENIEROS DE CAMINOS, CANALES Y PUERTOS

DEPARTAMENTO DE INGENIERÍA CIVIL

PROGRAMA DE DOCTORADO:
"Seguridad, Calidad y Optimización de Recursos en Infraestructuras y su Relación Medioambiental (242.99.2)"

TESIS DOCTORAL

ANÁLISIS DE LA CALIDAD DEL SERVICIO DEL TRANSPORTE PÚBLICO MEDIANTE ÁRBOLES DE DECISIÓN

ANALYSIS OF SERVICE QUALITY IN PUBLIC TRANSPORTATION USING DECISION TREES

Para la obtención del
GRADO DE DOCTOR POR LA UNIVERSIDAD DE GRANADA CON MENCIÓN DE DOCTORADO INTERNACIONAL

AUTOR:

ROCÍO DE OÑA LOPEZ

Granada, 2013

Editor: Editorial de la Universidad de Granada
Autor: Rocío de Oña López
D.L.: GR 102-2014
ISBN: 978-84-9028-679-1

UNIVERSIDAD DE GRANADA

E.T.S. INGENIEROS DE CAMINOS, CANALES Y PUERTOS

DEPARTAMENTO DE INGENIERÍA CIVIL

TESIS DOCTORAL

ANÁLISIS DE LA CALIDAD DEL SERVICIO DEL TRANSPORTE PÚBLICO MEDIANTE ÁRBOLES DE DECISIÓN

ANALYSIS OF SERVICE QUALITY IN PUBLIC TRANSPORTATION USING DECISION TREES

AUTOR:

ROCÍO DE OÑA LOPEZ

DIRECTOR:

JUAN DE OÑA LÓPEZ

Doctor Ingeniero de Caminos, Canales y Puertos

Universidad de Granada

TESIS DOCTORAL-2013

Memoria presentada por Rocío de Oña López para aspirar al grado de Doctor por la Universidad de Granada con mención de Doctorado Internacional

La doctoranda, Rocío de Oña López, y el director de la tesis "Analysis of Service Quality in Public Transportation using Decision Trees", garantizamos, al firmar esta tesis doctoral, que el trabajo ha sido realizado por el doctorando bajo la dirección de los directores de la tesis y hasta donde nuestro conocimiento alcanza, en la realización del trabajo, se han respetado los derechos de otros autores a ser citados, cuando se han utilizado sus resultados o publicaciones.

Granada, Marzo de 2013

Director/es de la Tesis

Fdo.: Juan de Oña López

Doctoranda

Fdo.: Rocío de Oña López

AGRADECIMIENTOS

Quisiera dedicar mis primeras palabras de gratitud a mi director de tesis y hermano D. Juan de Oña López, quién a pesar de la dificultad que conlleva ejercer ambas figuras a la vez, ha sabido separar en todo momento lo personal de lo profesional, siendo un excepcional director, asesor y guía de esta tesis, pero siendo aún más si se puede mejor hermano. Sin su confianza, apoyo, dedicación y cariño este trabajo no podría haberse llevado a cabo.

Mi agradecimiento también a todos los compañeros de la E.T.S. Ingenieros de Caminos, Canales y Puertos, y especialmente a mis compañeros del Área de Transportes, Laura, Gridelda y Pachi, con los que he compartido risas, lágrimas y emociones a lo largo de esta etapa. Gracias por vuestro apoyo incondicional y por hacer que el día a día en la Universidad sea tan agradable con vuestra presencia.

Gracias a mis padres, por su confianza e impulso en los momentos difíciles. A mis hermanos y demás familiares que demostraron siempre su apoyo y cariño. Y como no a Eduardo, mi marido y ante todo amigo, quién más que nadie ha estado conmigo durante todo este tiempo. Porque tienes el don de hacerme feliz. Gracias.

Quiero agradecer también a mis amigas y amigos por estar siempre ahí, y ser capaces de arrancarme una sonrisa bajo cualquier circunstancia. Todos, indirectamente, han puesto su granito de arena para la elaboración de esta tesis.

Tampoco puedo olvidarme de Dña. Gabriella Mazzulla y Dña. Laura Eboli, quienes me acogieron con gran afecto y cariño durante mi estancia en la Universidad de Calabria. Muchas gracias por darme la oportunidad de aprender de vosotras y trabajar con vosotras, además de permitirme desarrollar parte de este trabajo durante la estancia.

Por último agradecer al Consorcio de Transportes del área de Granada por proporcionarnos los datos utilizados en esta investigación. Igualmente, gracias a la Conserjería de Innovación,Ciencia y Economía de la Junta de Andalucía for la financiación del Proyecto de Excelencia denominado "Q-METROBUS-Quality of service indicator for METROpolitan public BUS transport services", al cual está asociada la beca que ha permitido que me dedique a la investigación.

RESUMEN

Hoy en día el éxito de un servicio de transporte público depende en gran medida del número de pasajeros que es capaz de atraer y retener. Por esta razón, la calidad del servicio se convierte en un aspecto de máxima importancia ya que una mejora en el nivel de calidad del servicio, provocará una mayor satisfacción de los pasajeros y un incremento en el uso del sistema.

Por lo tanto, actualmente, una de las principales preocupaciones de los planificadores del transporte es promocionar un servicio de transporte público de alta calidad. Con ello esperan disuadir la utilización del vehículo privado dentro de las ciudades y áreas metropolitanas a favor de una movilidad más sostenible.

Las técnicas que más se utilizán para analizar la calidad del servicio en el transporte público son aquellas basadas en encuestas de satisfacción, ya sea determinando un índice global de calidad o analizando los atributos del servicio por separado. Uno de los aspectos claves que debe de considerarse cuando se van a desarrollar índices para evaluar la calidad del servicio, es determinar cuánto peso dan los pasajeros a cada uno de los atributos cuando hacen su valoración global de la calidad. El método más utilizado por las empresas operadoras es pedir directamente a los pasajeros que puntuen la importancia de cada uno de los atributos en una escala determinada. Sin embargo, los métodos de importancias derivadas cuentan con gran cantidad de ventajas sobre los métodos de importancias declaradas. Estos métodos determinan el peso de cada uno de los atributos evaluando de forma estadística la fuerza de la relación existente entre estos atributos y la calidad global del servicio.

La mayor parte de las técnicas utilizadas para derivar la importancia de los atributos sobre la calidad percibida por los pasajeros tienen sus propias hipótesis y relaciones preestablecidas entre la variable dependiente y las independientes. Si estas hipótesis no se cumplen, el modelo podría realizar estimaciones erróneas.

Los Árboles de Decisión, es una técnica novedosa no paramétrica de minería de datos que no necesita que se cumpla ningún tipo de hipótesis, ni que existan relaciones predefinidas entre la variable dependiente y las independientes. Por este motivo, el principal objetivo de esta tesis doctoral es validar la utilización de esta metodología para analizar la calidad del servicio en el transporte público. Esta, además, se convierte en una herramienta potencial para los planificadores del transporte debido a la gran utilidad practica que proporcionan sus resultados, dada su simplicidad, su fácil interpretación, la posibilidad de extraer reglas, la habilidad para derivar la importancia de los atributos, etc.

Para validar la utilización de los árboles de decisión para analizar la calidad del servicio en el transporte público, se utilizaron los datos recogidos en varias encuestas de satisfacción en dos modos de transporte público distintos (un servicio de autobús metropolitano y un servicio de ferrocarril de cercanías) y en dos contextos diferentes (datos de España e Italia).

Además, se realizó un análisis detallado sobre la evaluación de la calidad del servicio entre grupos de usuarios para demostrar que las opiniones de los pasajeros eran heterogéneas entre ellos. Esta investigación evidencia la existente necesidad de analizar la calidad del servicio por grupos homogéneos de pasajeros, así como de formular estrategias de transporte personalizadas dirigidas a grupos específicos de usuarios.

Los resultados de la investigación mostraron que los Árboles de Decisión pueden ser utilizados de forma efectiva para analizar la calidad del servicio en el transporte público, ya que obtuvieron altos valores de precisión, e identificaron los factores claves que influencian la calidad global del servicio. Además, esta metodología cuenta con un valor añadido a las comúnmente empleadas, y es que extrae útiles reglas de decisión, en las que se explica la interacción de las variables que participan en el modelo.

ABSTRACT

Nowadays the success of a public transport system depends on the number of passengers which it is able to attract and retain. For this reason the quality of a service becomes an issue of maximum importance because it is known that an improvement in the level of quality of the service leads to a higher satisfaction of the passengers and to an increase in the use of the system.

Therefore, promoting high quality public transport services is one of the main current concerns of transport planners, who focusing in an on-going quality enhancement, seek to discourage the use of private cars in cities and metropolitan areas in favour of a more sustainable mobility.

The most popular techniques used for analyzing service quality in public transportation are those based on customer satisfaction surveys, arriving at a global index or analyzing the service attributes separately. A key aspect to take into consideration when developing indices to evaluate transit service quality is to determine how much weight passengers give to each attribute when making a global assessment of service quality. Asking customers to rate each attribute on an importance scale is the method mostly used by the operating companies. However, derived importance methods, which determinate the importance of the attribute by statistically testing the strength of the relationship of individual attributes with overall service quality, present a great number of benefits.

Most of these techniques used for derived the importance of the attributes on the passengers' perceived quality have their own model assumptions and pre-defined underlying relationships between dependent and independent variables. If these assumptions are violated, the model could lead to erroneous estimations of the likelihood of quality of service.

Decision Trees is a novel non parametric data mining technique which does not pre-define underlying relationship between the dependent variable and the independent variables. For this reason, the main objective of this Ph.D. thesis has been to validate that this methodology is appropriate for analyzing service quality in public transportation. Moreover, this methodology becomes a potencial tool for transport planners due to the great practical utility that provide the results, because of the simplicity, easiness of understanding the outcomes of the model, the posibility of extract rules, the ability of deriving the importance of the attributes, etc.

Data from various Customer Satisfaction Surveys collected in two different modes of public transport services (a metropolitan bus service and a suburban rail service) and two different contexts (data from Spain and Italy), were used for validating that Decision Trees is an appropiate methodology for analyzing service quality in public transportation.

Moreover, a detailed analysis about the evaluation of service quality among market segments was developed in order to prove that passengers opinions were heterogeneous. This research work evidence the neccesity of analyze service quality for more homogeneous groups of passengers and that transport planners should formulate personalized strategies (i.e. personalized marketing).

The results of this research work showed that Decision Trees can be used for effectively analyze service quality in public transport services, predicting service quality with high accuracy rates, and identifying the key factor influencing the overall service quality. Moreover, an added value of this technique was the useful decision rules extracted by the models, which explained the interaction of the variables participating in the model.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Chapter 1

INTRODUCTION

1.1. Statement of the problem

Promoting more sustainable modes of transport to alleviate the problems resulting from excessive use of the private car in most metropolitan areas (congestion, pollution, noise, etc.) is one of the main concerns of transport planners. Therefore, public transport service managers seek to diminish the use of private cars by promoting a consumer-based public transport service and on-going quality enhancement that will lead to higher customer satisfaction. According to the Handbook for Measuring Customer Satisfaction and Service Quality (HMSCCQ) (TRB, 1999), an increase in customer satisfaction translates into retained markets, increased use of the system, newly attracted customers, and a more positive public image.

Service Quality (SQ) is related to a series of attributes that describe the public transport service. To a large degree, it depends on the decisions that system managers adopt regarding the scope of the service (in terms of territory and schedules), the type of service provided, and so on. Many authors consider that SQ should be measured from the customer's perspective since, as Berry et al. (1990) point out, “customers are the sole judges of service quality”. Therefore, SQ can be measured by capturing passengers' perception of the attributes that describe the service.

Then, operating companies, in order to design appropriate transport strategies, every year or with a six-month frequency, monitor the perceptions of the users about the service. These perceptions are usually measured by Customer Satisfaction Surveys (CSS), and the data collected provide them useful information about different service quality characteristcs and their performance evolution along the time. However, in order to carry out more effective transport policies, they need not only to know the perceptions about the quality attributes, but also to identify which of these attributes have the highest influence on the global assessment of the service.

Several approaches have been used to estimate the relative importance of each attribute with regards to the SQ perceived by each customer. The methods can be classified as stated importance methods (asking customers to rate each attribute on an importance scale) or derived importance methods (deriving a measure of attribute importance by statistically testing the strength of the relationship of individual attributes with overall satisfaction). Although derived importance methods are preferred by researches because of their numerous advantages (Weinstein, 2000), they are not very used by public transport managers because of their higher complexity.

Many authors (Eboli and Mazzulla, 2008; 2011; Del Olio et al., 2010; Cirillo et al., 2011) point to the heterogeneity of passengers' perception about different aspects of the service. The heterogeneity is due to the qualitative nature of certain aspects that characterize the services, the different attitudes passengers have towards the use of PT, the different ways of viewing aspects of the service, and the social and economic characteristics of passengers and their preferences (Eboli and Mazzulla, 2011). It has even been shown that the same person may change his or her evaluation if they are made to reflect on certain important aspects of the service (Del Olio et al., 2010).

This heterogeneity represents a problem for many techniques that intend to measure SQ. Some authors (e.g. Dell'Olio et al., 2010) propose specific models after conducting stratified sampling based on the social and demographic characteristics of the passengers (i.e. models for women, for the elderly, according to income level, etc.). This poses two limitations, however: (a) if the samples are small, stratifying is a problem because a data set may be under-represented, and (b), it may be possible to obtain the weight of the variables entered in the model (service characteristics and perceptions), but the weight of the socio-economic characteristics and travel habits variables in the model are impossible to know. Other authors (Eboli and Mazzulla, 2008; Cirillo et al., 2011) have proposed mixed logit models to introduce such heterogeneity in the models.

However, most of these models have their own model assumptions and pre-defined underlying relationships between dependent and independent variables. If these assumptions are violated, the model could lead to erroneous estimations of the likelihood of quality of service.

The decision trees, and particularly the classification and regression tree (CART), a non-parametric model with no pre-defined underlying relationship between the target (dependent) variable and the predictors (independent variables), has been widely employed in business administration, agriculture, industry, and engineering. With the ability to automatically search for the best predictors and the best threshold values for all predictors to classify the target variable, CART has been shown to be a powerful tool, particularly for dealing with prediction and classification problems.

Therefore, the main purpose of this study is to examine whether or not the CART methodology can effectively analyze service quality in public transport services and identify the key factors affecting it.

1.2. Objectives

The main objective of this Ph.D. thesis is to validate that Decision Trees is an appropiate methodology for analyzing service quality in public transportation. Decision Trees is a novel statistic technique for analyzing service quality in public transportation and it presents a high practical utility for transport managers due to its simplicity, its easiness of understanding, the possibility of formulate rules, the ability for deriving the importance of the attributes, and so on. In order to comply this major objective a set of specific objetives are also proposed in this research work, such as identifying the most relevant variables influencing the overall service quality, demostrate that passengers opinions change before and after they are made to reflect on the attributes describing the service, validate that the key factors influencing the overall SQ are different among market segments, inquire in the problematic of stated importance rates about SQ attributes and research if the passengers' socioeconomic characteristics and travel habits variables are influent in the overall evaluation about the service.

By verifying all of these objectives a better undertanding about service quality could be provided to transport managers and operators, permitting them to design adequate marketing policies that promote increasing the use of public transport services and therefore, a more sustainable mobility.

1.3. Thesis organization

In this section a brief and concise description of the estructure of this Ph.D. thesis is carried out. This thesis consists in seven chapters:

Chapter 1 includes an introduction to the thesis, a brief description of the proposed objectives, the structure of the document and the main contributions of this research.

Chapter 2 presents an overview of the general characteristics of service quality in the public transport sector, the methodological issues associated with its analysis, a discussion of the main methodological approaches used, and a brief introduction to Decision Trees, their main applications, advantages and disadvantages.

Chapter 3 presents the objectives to be fulfilled in this research work.

Chapter 4 presents the methodology and datasets used in this thesis. A description of the methodology followed in the thesis, the algorithm used to build the Decision Trees, the techniques used to validate and evaluate the model, the algorithm followed to extract the importance of the variables and the generation of the decision rules is performed in the first part of the chapter. The datasets used for building the models in two different experimental contexts are also described in this chapter.

Chapter 5 shows the results of modeling service quality in two different public transport services (a bus metropolitan transit service and a suburban rail service) using Decision Trees.

Chapter 6 presents the major conclusions of this work and future research lines.

Finally, **Chapter 7** includes all the references used in this thesis.

1.4. Main Contributions

The main contributions of this thesis are four papers, that have been published or they are under review, and four articles in International Conferences. They represent the main findings obtained by this research work. The mentioned contributions are the following:

Paper 1. Juan de Oña, Rocío de Oña, F. Javier Calvo. A classification tree approach to identify key factors of transit service quality. Published in *Expert Systems with Applications, 39, 11164-11171*

Paper 2. Juan de Oña and Rocío de Oña. Quality of service in public transport based on customer satisfaction surveys: A review and assessment of methodological approaches. Paper under Review in *Transpor Reviews*

Paper 3. Rocío de Oña and Juan de Oña. Analysis of transit quality of service through segmentation and classification tree techniques. Paper under review in *Transport Policy*

Paper 4. Juan de Oña, Rocío de Oña, Laura Eboli and Gabriella Mazzulla. Heterogeneity in perceptions of service quality among groups of railway passengers. Paper under Review in *International Journal of Sustainable Transportation*

Conferece 1. Rocío de Oña, Laura Eboli and Gabriella Mazzulla (2012). Key factors affecting rail service quality. A decision tree approach*. XIX Conference SIDT,* Padua (Italy), 18-19 October, 2012.

Conference 2. Rocío de Oña and Juan de Oña. Using decision trees for analyzing behavioural intentions in transit service quality. *MAMERN' 2013, International conference on approximation methods and numerical modelling in environment and natural resources,* Granada (Spain), 22-25 April, 2013

Conference 3. Rocío de Oña and Juan de Oña. Analyzing Transit Service Quality Evolution Using Decision Trees and Gender Segmentation. *19th International Conference on Urban Transport and the Environment,* Kos (Greece), 29-31 May, 2013. Accepted for oral presentation.

Conference 4. Rocío de Oña, Laura Eboli and Gabriella Mazzulla. Monitoring changes in transit service quality over time. *16th Euro Working Group on Transportation*, Oporto (Portugal), 4-6 september, 2013. Under Review

CHAPTER 2

STATE OF THE ART

Chapter 2

STATE OF THE ART

This chapter provides a review of contemporary thinking on Public Transport quality of service-analysis field and highlights the main methodological approaches that have been used to address this issue. To this end, the general characteristics of Service Quality (SQ) in the public transport sector and methodological issues associated with its analysis are discussed. Then, a critical assessment of the various methodological approaches that have been used to analyze SQ in the public transport sector has been carried out.

Finally, it concludes with a brief description and review of existing literature about a novel methodology for analyzing SQ, which emerged in the last years among other widespread techniques as a powerful tool in other research fields and that could be applied for analyzing SQ in the public transport sector in the future.

2.1. Quality of service in public transportation

For a long time the performance evaluation of public transport has been carried out from the service managers' perspective (transport company and government), based on the cost efficency and cost effectivenes of public transport services and operations (Carter and Lomax, 1992; Fielding, 1992; Fielding et al., 1985; Hensher and Daniels, 1995; Ozment and Morash, 1998; Pullen, 1993; Wipper, 1993). Little attention was paid to the point of view of passengers. However, in the last few decades, SQ has become a major area of attention for practitioners, managers and researchers, who have focused on the passengers' perspective.

Currently, researchers and managers in the public transport sector strive for learning details about the main factors affecting SQ in their organizations for the obvious reasons of customer satisfaction, increased profitability, etc. In this context, models gain specific importance as they not only help to learn the factors associated with SQ but also provide a direction for improvements.

Many authors have studied SQ in the public transport sector from varying perspectives and using different methodologies in recent years. The variety of existing approaches could be justified by the complexity of the service quality concept; the number of attributes used to evaluate it; the imprecision and subjectivity of the data used to analyse it, typically based on Customer Satisfaction Surveys (CSS); and the heterogenity of passenger perceptions of public transportation.

The beginning of the 21st century saw an increase in the use of discrete choice models based on Stated Preference surveys (Hensher, 2000; 2001; Hensher and Prioni, 2002; Hensher et al., 2003) to analyze public transport SQ. Such methods are based on the assumption that although specific aspects of SQ may be particularly positive or negative in a passenger's satisfaction with a service, the overall level of passenger satisfaction is best measured by how an individual evaluates the total package of services on offer (Hensher and Prioni, 2002). Nonetheless, models based on CSS have been and are still the most widely adopted models for analyzing SQ in the public transport sector (see Table 1). So, we have focused this state of the art review on this kind of surveys and on the models that use them.

2.1.1. General characteristics of service quality in public transport and methodological issues

Past research has identified a number of characteristics and methodological issues that are critical considerations in the development and application of an appropriate methodology to analyze the service quality in public transport. A summary of these characteristics and methodological issues is presented in this section.

Table 1. Summary of previous research: Public transport industry, scope of study, valid surveys and scale used

References	Public transport industry	Scope of study	Valid surveys	Scale used
Airlines and Airports				
Abdlla et al., 2007	Airlines	Egypt	474	9-point Likert
Aksoy et al., 2003	Airlines	Istanbul Airport (Turkey)	1.014	7-point Likert
Chang and Yeh, 2002	Airlines	Taiwan	354	11-point scale
Chau and Kao, 2009	Airlines	Taipei (Taiwan) and London (UK)	161 and 102	5-point Likert
Chen and Chang, 2005	Airlines	Taiwan	470	5-point Likert
Chen, 2008	Airlines	Taiwan	245	5- & 7-point Likert
Cheng et al., 2008	Airlines	Taiwan	252	5-point Likert
Chou et al., 2011a	Airlines	Taiwan	329	5-point linguistic
Forgas et al., 2010	Airlines	Barcelona-London corridor	1.700	n.a.
Gilbert and Wong, 2003	Airlines	Hong Kong Airport	365	8-point scale
Huse and Evangelho, 2007	Airlines	Santos Dumont Airport (Brasil)	88	10-point Likert
Kiatcharoenpol and Laosirihongthong. 2006	Airlines	Developing country	n.a.	5-point Likert
Kim and Lee, 2011	Airlines	South Korea	244	5-point Likert
Kim et al., 2011	Airlines	South Korea	231	5-point Likert
Kuo, 2011	Airlines	China-Taiwan corridor	1.635	7-point linguistic
Liou and Tzeng, 2007	Airlines	Taiwan	408	11-point scale
Liou et al., 2011b	Airlines	Taiwan	5.553	5-point Likert
López-Bonilla and López-Bonilla, 2008	Airlines	Spain	3,000 and 1,911	5-point Likert
Nejati et al., 2009	Airlines	Teheran (Iran)	231	7-point Likert
Ostrowski et al., 1993	Airlines	USA	6.000	4-point scale
Oyewole, 2001	Airlines	n.a.	261	10-point Likert
Pakdil and Aydin, 2007	Airlines	Turkey	298	5-point Likert
Park et al, 2004	Airlines	Korea	592	7-point Likert
Park et al., 2006	Airlines	Sydney Airport (Australia)	501	7-point Likert
Park, 2007	Airlines	Incheon International Airport (Korea) and Sydney Airport (Australia)	592 and 501	7-point Likert
Ringle et al., 2011	Airlines	International Airport (Western Europe)	1.031	n.a.
Ritchie et al., 1980	Airlines	Calgary (Canada)	150	7-point Likert
Saha and Theingi, 2009	Airlines	Thailand	1.212	7-point Likert
Sultan and Simpson, 2000	Airlines	North Transaltantic corridor	1.956	7-point Likert
Surovitskikh and Lubbe, 2008	Airlines	Middle Eastern Airlines in South Africa	410	7-point Likert
Tsaur et al., 2002	Airlines	Taiwan	211	5-point linguistic
Wen et al., 2008	Airlines	Taipei-Tokio corridor	381	7-point Likert

Yang et al., 2012	Airlines	Taiwan	458	5-point Likert
Fernandes and Pacheco, 2010	Airports	Brazil	947	3-point linguistic
Kuo and Liang, 2011	Airports	Northeast-Asian region	23 and 26	7-point linguistic
Liou et al., 2011a	Airports	Taoyuan International Airport (Taiwan)	503	3-point Likert
Tsai et al., 2011	Airports	Taoyuan International Airport (Taiwan)	204	n.a.
Yeh and Kuo, 2003	Airports	Taiwan	15	5-point Likert
Urban and Metropolitan Public Transport				
Andreassen, 1995	Bus and rail services	Oslo Area (Norway)	1.000	7-point Likert
Christopher et al., 1999	Bus and rail services	Chicago (USA)	>2,400	5-point Likert
Foote et al., 2001	Bus and rail services	Chicago (USA)	2.464	5-point Likert
Karlaftis et al., 2001	Bus and rail services	Athens (Greece)	n.a.	n.a.
Minser and Webb, 2010	Bus and rail services	Chicago (USA)	264	5-point Likert
Tyrinopoulos and Antoniou, 2008	Bus and rail services	Athens and Thessaloniki (Greece)	1.474	4- & 5-point Likert
Figler et al., 2011	Bus services	Chicago (USA)	364	5-point Likert
Foote and Stuart, 1998	Bus services	Chicago (USA)	4.191	5-point Likert & 11-point scale
Friman, 2004	Bus services	Sweden	2.797	9-point Likert
Glascock, 1997	Bus services	Seattle (USA)	485	n.a.
Hensher et al., 2010	Bus services	Tyne and Wear area (UK)	310	5-point Likert
Jen and Hu, 2003	Bus services	Taipei (Taiwan)	235	5-point Likert
Koushki et al., 2003	Bus services	Kuwait	679	5-point Likert
Diana, 2012	Bus, trolley and tram services	Urban areas (Italy)	4,123	4-point scale
Eboli and Mazzulla, 2007	Metropolitan bus services	University of Calabria, Cosenza (Italy)	763 (students)	10-point Likert
Eboli and Mazzulla, 2009	Metropolitan bus services	Cosenza, Calabria (Italy)	218	10-point Likert
Eboli and Mazzulla, 2011	Metropolitan bus services	Cosenza and Rende (Italy)	123	11-point scale
Hu, 2010	Metropolitan bus services	Taipei (Taiwan)	292	7-point Likert
Friman and Gärling, 2001	PT services	Sweden	95	Number from 10 (very dissatisfied) to 90 (very satisfied)
Friman et al., 2001	PT services	Sweden	997	9-point Likert
Pedersen et al., 2011	PT services	Stockholm (Sweden)	1,007 and 169	5-point Likert & 11-point scale
Awasthi et al., 2011	Railways (subway)	Montreal (Canada)	60	5-point Likert
Lai and Chen, 2011	Railways (subway)	Kaohsiung (Taiwan)	763	5-point Likert
Stuart et al., 2000	Railways (subway)	New York (USA)	1.075	11-point scale
Weinstein, 2000	Rapid-transit system (trains)	Bay Area District, San Francisco (USA)	4.150	5-point linguistic & 7-point Likert
Dell'Olio et al., 2010	Urban bus services	Santander (Spain)	768	5-point Likert

Hu and Jen, 2006	Urban bus services	Taipei (Taiwan)	3 data collection (1: 244; 2: 292; 3: 235)	7-point Likert
Sánchez et al., 2007	Urban bus services	Almeria (Spain)	1.000	5-point Likert
Yeh et al., 2000;	Urban bus services	Taipei (Taiwan)	n.a.	5-point linguistic
Castillo and Benitez, 2012	Urban bus services	Bilbao (Spain)	1.508	11-point scale
Murray et al., 2010	Urban public services	Auckland, Wellington and Christchurch (New Zealand)	639	5- & 7-point Likert
Nurul-Habib et al., 2011	Urban public services	Calgary (Canada)	500	5-point Likert
Wang et al., 2010	Urban public services	Taipei Metropolitan Area	510 and 103	5-point Likert
Interurban Public Transport				
Jen et al., 2011	Interurban bus services	Taiwan	747	7-point Likert
Kuo et al., 2007	Interurban bus services	Taiwan	60	7-point linguistic
Lin et al., 2008	Interurban bus services	Taiwan	385	5-point Likert
Wen et al., 2005	Interurban bus services	Taiwan	600	5-point Likert
Cavana et al., 2007	Railways	Wellington (New Zealand)	340	9-point Likert
Drea and Hanna, 2000	Railways	USA	2.369	n.a.
Ganesan-Lim et al., 2008	Railways	Queensland (Australia)	224	7-point Likert
Liu and Gao, 2007	Railways	China	168	n.a.
Nathanail, 2008	Railways	Greece	n.a.	5-point Likert
Tripp and Drea, 2002	Railways	Illinois (USA)	2.529	7-point Likert
Chou and Kim, 2009 ; Chou et al., 2011b	Railways, high speed	Taiwan and Korea	418 and 414	10-point Likert
Others				
Paquette et al., 2012	Dial-a-ride services	Montreal (Canada)	331	10-point Likert
Mathisen and Solvoll, 2010	Ferry passenger	Norway	1.734	5-point Likert
Joewono and Kubota, 2007a ; 2007b ; 2007c	Paratransit	Indonesia	980	5-point Likert
Stradling et al., 2007	Three different transport industries	Scotland (UK)	213, 666 and 1,101	5-point Likert

n.a.: not available

[a] This is a representative, but not a comprehensive list of references.

2.1.1.1. Complexity of the quality concept

The concept of SQ is complex, fuzzy and abstract, mainly because of the three properties of service: intangibility, heterogeneity and inseparability (Carman, 1990; Parasuraman et al., 1985):

- Intangibility, in that their outputs cannot be measured in terms of their physical attributes (services are performances rather than objects and are experienced by the customer).

- Heterogeneity, in that the service is likely to be different for each individual who receives it.

- Inseparability of production and consumption (services are sold and then produced and consumed at one and the same time).

An important number of authors (Grönroos, 1988; Lehtinen and Lehtinen, 1982; Parasuraman et al., 1985; Sasser et al., 1978) maintain that the perception of SQ is the result of a comparison of consumer expectations with actual service performance perception. Other authors, however, do not take expectations into consideration (Cronin and Taylor, 1992). They are only interested in passengers' perceptions, or even the perception of transport companies and government managers (Eboli and Mazzulla, 2011; 2012; Nathanail, 2008; Tyrinopoulos and Aifadopoulou, 2008).

There is no consensus on consumer expectations. Certain models in the literature compare customer performance perception with ideal performance or quality (Lin et al., 2008; Mattsson, 1992); with desired quality (Cavana et al., 2007; Gilbert and Wong, 2003); and with adequate or tolerable quality (Hu and Jen, 2006). Teas (1993) stated that expectations could be interpreted as predictions of service, as an ideal standard or as attribute importance. When analysing SQ in the public transport sector, a number of researchers (Abdalla et al., 2007; Aksoy et al., 2003; Chau and Kao, 2009; Chen and Chang, 2005; Eboli and Mazzulla, 2010; Gilbert and Wong, 2003; Sultan and Simpson, 2000) have substituted importance measures for expectations, although there is no theoretical basis for this (Landrum and Prybutok, 2004) and importance ratings differ from the expectation ones (Smith, 1995). Landrum and Prybutok (2004) found that importance and expectations are not the same construct, but they indicated that comparing service performance against what customers consider important may be just as useful to managers as comparing performance against what customers expect. Considering the variety of ways that expectations can be interpreted, importance ratings may be less conducive to confusion. Also, from the point of view of Smith (1995), measuring which service attributes are important to customers may be more meaningful to managers than measuring customer service expectations.

The relationship between SQ and satisfaction is not clear. In the literature, SQ usually accompanies satisfaction. This may be due to the similar nature of the two variables, which both derive from the disconfirmation theory (Parasuraman et al. 1988). Some authors are of the opinion that customer satisfaction causes perceived quality and others consider that SQ is a vehicle for customer satisfaction (Chen, 2008; Chou and Kim, 2009; Chou et al., 2011b; Cronin and Taylor, 1992). In recent years, a lot has been said about the “Service Quality–Satisfaction–Loyalty/Behavioural Intentions” paradigm (Choi et al., 2004; Cronin Jr et al., 2000; Dabholkar et al., 2000; Fornell et al., 1996; González et al., 2007; Jen et al., 2011; Ledden et al., 2007). This paradigm suggests that satisfaction is the link between SQ and loyalty or behavioural intentions. Therefore, it would be on a “higher” attitude level with regards to SQ (Mattsson, 1992). In the models of Mattsson (1992) and of Spreng and Mackoy (1996), satisfaction is reached by comparing expectation and performance. Thus, SQ and satisfaction are used interchangeably in much of the literature (Cavana et al., 2007).

Grönroos (1982, 1984, 1988) and Lehtinen and Lehtinen (1991) support the three-dimensionality of SQ in terms of technical quality (the quality of what consumer actually receives), functional quality (how he gets the technical outcome) and image (the image attributed to service providers by their current and potential consumers). However, functional quality seems to be more important than technical quality (Grönroos, 1984). Parasuraman et al. (1985, 1988) pointed out that service offers very few tangible elements, and therefore they focused their efforts on intangible elements (functional quality).

2.1.1.2. Attributes

A very large number of attributes are used to evaluate SQ[1], so they are normally grouped into a smaller number, called dimensions.

Although there is no general agreement as to the nature or content of SQ dimensions (Brady and Cronin Jr, 2001), there is a general recognition that service quality is a multidimensional construct (Brady and Cronin Jr, 2001; Cronin and Taylor, 1992; Parasuraman et al., 1985; 1988), and multilevel or hierarchical (Dabholkar et al, 1996; Brady and Cronin Jr, 2001; Jen et al., 2011).

Various papers (e.g. Eboli and Mazzulla, 2008; Philip and Hazlett, 1997; TRB, 2004; Tripp and Drea, 2002; UNE, 2003) have pointed to the existence of several categories of attributes that have a greater or lesser impact on SQ and satisfaction. Philip and Hazlett (1997) propose a model with a hierarchical structure, based on three classes of attributes: pivotal, core and peripheral attributes. This model was subsequently contrasted for the rail transportation industry by Tripp and Drea (2002). The pivotal attributes exert the greatest influence on the satisfaction levels. Core attributes are the amalgamation of the people, processes and the

[1]The UNE-EN 13816 (2003) standard considers 117 attributes that define a public transport service at the first level; at the second level, the variables are grouped into 30 sub-dimensions, and at the third level into 8 dimensions. Murray et al (2010) use 166 items on their survey to analyze the attitudes towards public transport in New Zealand.

service organizational structure through which consumers must interact and/or negotiate so that they can achieve/receive the pivotal attribute. And the peripheral attributes can be defined as the "incidental extras" designed to add "roundness" to the service encounter and make the whole experience for the consumer a complete delight. The UNE-EN 13186 (2003) standard classifies the service's characteristics into basic, proportional and attractive, depending on how compliance and non-compliance affects customer satisfaction (see Table 2).

Table 2. Service's characteristics classification by the UNE-EN 13186

Characteristics	Effects of non-compliance	Effects of compliance
Basic (they are attributes or prerequisites, commonly "expected" and considered implicit in the transport service concept) e.g. safety of the vehicle, punctuality,…	Little non-compliances poduce high insatisfaction (e.g. a delay in the service turns into a non reliable service)	Comply them in an adequate way does not have a great effect in the costumer's satisfaction (e.g. arrived punctuality it is considered as "fair")
Proportional (they are the normal attributes, which characterizise the denominated "quality" in the compliance of the service) e.g. confort, cleanliness, etc.	They cause a grade of insatisfaction proportional to the grade of non-compliance	The satisfaction is proportional to the grade of compliance (e.g. the costumer expects quality with the minor price and his satisfaction will be higher if he perceives that this relationship is adequate)
Attractive (they are interesting characteristics, distinguising the service from others, surprises that delight passengers)	The consumer does not have them in mind, then he/she does not evaluate their non-compliance (their absence will not produce insatisfaction)	Customers surprise with positive characteristics can increase their satisfaction (if the proportional characteristics are reasonably complied and the basic characteristics are totally complied). Their presence leads to high perceptions of the quality.

The Transit Capacity and Quality of Service Manual (TRB, 2004) groups attributes into availiablity factors (more important to passengers), and comfort and convenience factors (less important).

- *Availability Factors*. When the service is not available, others aspects of transit service quality will not matter to the passengers for that trip. These factors are the spatial availability, temporal availability, information and capacity.

- *Comfort and Convenience Factors.* When all the factors listed above are met, then transit becomes an option for a given trip. At this point, passengers weight the comfort and convenience of transit against competing modes. Some of the things that a potencial passenger may consider are: the load of passengers, reliability, time making the trip, security, cost, appearance and comfort.

Eboli and Mazzulla (2008) demonstrated the existence of two cathegories of attributes (basic and non-basic) empirically from the preferences showed by users. Basic attributes compromise SQ when their level is low, and non-basic attributes are considered secondary service characteristics that affect SQ if they are present, but do not compromise it if they are absent.

Parasuraman et al. (1988) defended the existence of a generic list of attributes and dimensions to analyse the SQ of any type of service. They defined a group of attributes and dimensions at the same time that the SERVQUAL scale (SERVice QUALity), that predominate in the most part of the quality studies. The authors, for preparing the questionnaires, carried out a series of interviews with different executives of enterprises and key customers of diverse industries: banks, hotel business, informatics, etc, with the intention of develop this questionnaire for all type of services.

However, many authors criticized the generic list (Babakus and Boller, 1992; Caro and Garcia, 2007; Jabnoun and Khalifa, 2005; Van Dyke et al., 1997). Carrillat et al. (2007) demonstrated that the predictive value of the Parasuraman et al. (1988) model increased when the model's items were adapted to the study context. This was because not all services shared the same characteristics, and therefore, the survey was not adjusted to the various needs required for different types of services.

Most authors agree that the attributes included in a survey must be selected to each specific case (Babakus and Boller, 1992; Brown et al., 1993; Carman, 1990; Van Dyke et al., 1997). The aspects appreciated by each user are highly dependent on the users' social and demographic characteristics (age, gender, occupation, education, marital status, household income, etc); their context (i.e. geographical area, social class and type of service); the reason for travel; and the modes of transport used (Andreassen, 1995; Ganesan-Lim et al., 2008; Gilbert and Wong, 2003; Oyewole, 2001).

The selection is frequently made on the basis of an exhaustive study of which attributes are the most important in terms of evaluating SQ in the service under study. Several methods are used to that end in the field of public transportation: literature review (Joewono and Kubota, 2007a; 2007b; 2007c), survey of operators

(Andreassen, 1995), focus groups[2] (Weinstein,2000), pilot users survey (Liou et al., 2011a), Churchill's paradigm (Churchill, 1979; Brown et al., 1993), statistical tests to identify whether an attribute should or should not be considered (i.e. confirmatory factor analysis, Cronbach's alpha, etc.). In most cases, combinations of these methods are used (e.g. Cavana et al., 2007; Chau and Kao, 2009; Chen and Chang, 2005; Dell'Olio et al., 2010; Gatta and Marcucci, 2007; Hensher and Prioni, 2002; Hensher et al., 2003; Hu and Jen, 2006; Liu and Gao, 2007; Pakdil and Aydin, 2007; Tyrinopoulos and Aifadopoulou, 2008). These methods are also used to simplify data collection by lowering the number of attributes.

2.1.1.3. Nature of the data

In passenger transport services, functional quality is more important than technical quality (Grönroos, 1984; Parasuraman et al., 1985; 1988) which gives the SQ concept a subjective nature, insofar as it is the result of passenger perceptions or its comparison with their expectations. Therefore, the evaluation process usually involves subjective assessments, resulting in qualitative and imprecise data being used.

Several authors (Awasthi et al., 2011; Chang and Yeh, 2002; Chang et al., 2012; Chou et al., 2011a; Fernandes and Pacheco, 2010; Kuo and Liang, 2012; Kuo, 2011; Nejati et al., 2009; Yeh and Kuo, 2003; Yeh et al. 2000) have proposed using the fuzzy set theory (Zadeh,1965) as an effective method for handling the issue of subjective, qualitative and imprecise information inherent in the data used to assess service quality.

In recent years, there has been an emerging debate on whether subjective data (customers' opinions) on SQ can be combined with objective data (technical data) on service performance to evaluate the global quality of public transport. Some authors (Parasuraman et al., 1988; Carman, 1990; Gourdin and Kloppenborg, 1991; Transportation Research Board, 2004) dismiss this approach because they consider that SQ is the quality perceived by the passengers' point of view. However, in the past ten years several studies have begun to propose the combined use of subjective and objective measures to determine SQ (Eboli and Mazzulla, 2011; 2012; Nathanail, 2008; Tyrinopoulos and Aifadopoulou, 2008; Yeh et al. 2000).

[2]Fowler (1995) stated that the focus group interview could improve a questionnaire in two ways. First, the related hypothesis regarding the investigation context can be examined. Second, it aids in evaluating the expressions in the questionnaires and the hypothesis regarding the language used or cognitive assumptions.

2.1.1.4. Surveys

User surveys are an essential tool for collecting the information used to analyse quality. As indicated previously, Customer Satisfaction Surveys (CSS) are widely adopted to analyse public transport quality, although the number of stated preference surveys is increasing during the last years, mainly among academics. CSS are questionnaires where, at least, customers are asked to rate satisfaction or performance perception on each key service attribute. In addition, customers are normally asked to answer other questions as well, depending on the methodological approach used for the subsequent data analysis (see Section 2.1.2). They are often asked to rate also the importance of each attribute (see Table 3) and global overall service satisfaction (Aksoy et al., 2003; Andreassen, 1995; Cavana et al., 2007; Chau and Kao, 2009; Christopher et al., 1999; Dell'Olio et al., 2010; Figler et al., 2011; Foote and Stuart, 1998; Foote et al., 2001; Friman and Gärling, 2001; Friman et al., 2001; Friman, 2004; Hensher et al., 2010; Joewono and Kubota, 2007a; 2007b; 2007c; Koushki et al., 2003; Liou et al., 2011a; Pakdil and Aydin, 2007; Park et al, 2004; Pedersen et al., 2011; Tyrinopoulos and Aifadopoulou, 2008; Tyrinopoulos and Antoniou, 2008; Weinstein, 2000). In some cases, they are asked for a rate on each attribute, in terms of both perceptions and expectations (Cavana et al., 2007; Chou et al., 2011a; Hu and Jen, 2006; Hu, 2010; Lin et al., 2008; Liu and Gao, 2007; Pakdil and Aydin, 2007; Park et al, 2004; Sultan and Simpson, 2000; Wang et al., 2010); or a rate on global service, in terms of both perceptions and expectations (Eboli and Mazzulla, 2007).

Normally, the rates are expressed on two scales: numeric or linguistic. Numeric scales are more widely used and have a wider range: from 3- to 11-points. Table 1 shows that the 5-point Likert scales are the most widely adopted. Linguistic scales are used less and have a narrower range: from 3- to 7-points.

2.1.1.5. Heterogeneity

The dimensions of quality, viewed from a customer's perspective, are complex, and perceptions about qualitative characteristics of service are very different among users. Users' perceptions of public transport services are heterogeneous for many reasons: the qualitative nature of some service aspects, the different users' socioeconomic characteristics, and the diversity in tastes and attitudes towards public transport.

To analyse this heterogeneity, one possibility is to stratify the sample and then build specific models. Segmentation is normally carried out in terms of the survey population's socioeconomic and demographic characteristics (i.e. income, gender, age, car availability, frequency, etc.) (e.g. Andreassen, 1995; Dell'Olio et al., 2010). However, other procedures are also used, such as cluster analysis (e.g. Wen et al., 2008).

Table 3. Summary of previous research on public transport classified by the methods used to determine the weight of the attributes used in SQ analyses

TECHNIQUE	PREVIOUS RESEARCH IN PUBLIC TRANSPORT
Asking for importance directly through Customer Satisfaction Surveys	
Abdlla et al., 2007; Aksoy et al., 2003; Awasthi et al., 2011; Cavana et al., 2007; Chang and Yeh, 2002; Chen and Chang, 2005; Chou et al., 2011a; Christopher et al., 1999; Eboli and Mazzulla, 2007; 2009; 2011; Fernandes and Pacheco, 2010; Foote and Stuart, 1998; Gilbert and Wong, 2003; Glascock, 1997; Hensher et al., 2010;Hu, 2010; Huse and Evangelho, 2007; Kuo and Liang, 2012; Kuo, 2011; Liou et al., 2011b; Liou and Tzeng, 2007; Mathisen and Solvoll, 2010; Nathanail, 2008; Nejati et al., 2009; Ostrowski et al., 1993; Paquette et al., 2012; Ritchie et al., 1980; Sánchez et al., 2007; Stradling et al., 2007; Sultan and Simpson, 2000; Surovitskikh and Lubbe, 2008; Tsai et al., 2011; Tsaur et al., 2002; Tyrinopoulos and Aifadopoulou, 2008; Tyrinopoulos and Antoniou, 2008; Wang et al., 2010; Wen et al., 2008; Yeh and Kuo, 2003; Yeh et al., 2000	
Model deduction from Customer Satisfaction Surveys	
Bivariate Pearson correlations	Figler et al., 2011; Weinstein, 2000
Regression analysis	***Multiple Linear Regression***: Kim and Lee, 2011; Weinstein, 2000 ***Generalized Linear Model:*** Castillo and Benitez, 2012 ***Discriminant Analysis (DV is categorical)***: Aksoy et al., 2003 ***Ordered Logit (logistic distribution)***: Tyrinopoulos and Aifadopoulou, 2008;Tyrinopoulos and Antoniou, 2008 ***Ordered Probit (normal distribution)***: Dell'Olio et al., 2010; Huse and Evangelho, 2007 ***Generalized Ordered Logit (account heterogeneity)***: Hensher et al., 2010
Structural Equation Model	Andreassen, 1995; Chen, 2008; Cheng et al., 2008; Chou and Kim, 2009; Chou et al., 2011b; Eboli and Mazzulla, 2007; Friman and Gärling, 2001; Friman et al., 2001; Jen et al., 2011;Joewono and Kubota, 2007a; 2007c; Karlaftis et al., 2001; Kim and Lee, 2011; Lai and Chen, 2011; Minser and Webb, 2010; Nurul-Habib et al., 2011; Park et al., 2006; Saha

	and Theingi, 2009; Stuart et al., 2000; Tripp and Drea, 2002; Wen et al., 2005; Yang et al., 2012
Path analysis	Forgas et al., 2010; Jen and Hu, 2003; Joewono and Kubota, 2007b; Lin et al., 2008; Park et al., 2004; Ringle et al., 2011

2.1.2. Methodological approaches based on customer satisfaction surveys

There are two main theoretical approaches: (a) performance perception and expectations approach (Parasuraman et al., 1985); and (b) only performance perception approach (Cronin and Taylor, 1992). Moreover, there are also two types of methodological approaches, depending on whether SQ is measured by disaggregation (i.e. service attributes are analysed individually) or aggregation (when an aggregated analysis of attributes is used to obtain an overall Service Quality Index, SQI, or a Customer Satisfaction Index, CSI. In an aggregated analysis, it is essential to know the weight or importance of each attribute in terms of global quality in order to construct SQI. The manners in which the weights can be obtained are approached in Section 2.1.3.

In some cases the two approaches are used together to profit from the benefits of both. For instance, disaggregated models help to set priorities for service improvements. They help managers to choose from among a long list of service attributes to more optimally focus their organization's attention and resources (Weinstein, 2000). The models that provide a SQI (which are more widely used, as shown in Table 4) permit service to be analysed over time and to compare different services (e.g. territorial scope, suppliers, etc.).

2.1.2.1.- Aggregated performance-expectation models

The best known and most widely applied technique was proposed by Parasuraman et al. (1985). They proposed that SQ is a function of the differences between expectation and performance, from customer point of view. They developed a model based on gap analysis and defined the overall service quality as a function of perception and expectations, and it was defined as:

$$\boldsymbol{SQ} = \sum_{j=1}^{k} \left(\boldsymbol{P_{ij}} - \boldsymbol{E_{ij}}\right) \tag{1}$$

where k is the number of attributes; P_{ij} is performance perception of stimulus i with respect to attribute j; and E_{ij} is service quality expectation for attribute j that is the relevant norm for stimulus i.

Parasuraman et al. (1988) developed the SERVQUAL (SERVice QUALity) scale for measuring customers' perception of SQ. The original 10-dimension scale collapsed into 5-dimensions: reliability, responsiveness, tangibles, assurance and empathy, which capture functional quality. Later SERVQUAL's revisions (Parasuraman et al., 1991a; 1994) reduced the total number of items to 21, but the five dimensional structure remained the same. A number of authors have used the SERVQUAL scale for analyzing airline SQ (Abdlla et al., 2007; Chau and Kao, 2009; Kiatcharoenpol and Laosirihongthong, 2006; Sultan and Simpson, 2000) and Liu and Gao (2007) adapted the SERVQUAL scale for evaluating the railway transport service in China.

Table 4. Summary of previous research on Public Transport analyzing SQ by model type

	Performance perceptions and expectations [c]	Only performance perceptions
Disaggregated models		
No Importance	Cavana et al., 2007; Chang et al., 2012; Hu and Jen, 2006	
With Importance	Hu, 2010; Mathisen and Solvoll, 2010; Tsai et al., 2011	Chen and Chang, 2005; Christopher et al., 1999; Chou et al., 2011b; Eboli and Mazzulla, 2011; Figler et al., 2011; Foote and Stuart, 1998; Stradling et al., 2007; Weinstein, 2000
Aggregated models [b]		
	Abdlla et al., 2007; Chang and Yeh, 2002; Chau and Kao, 2009; Chou et al., 2011a; Eboli and Mazzulla, 2009; Kiatcharoenpol and Laosirihongthong. 2006; Kuo and Liang, 2011; Kuo et al., 2007; Liou and Tzeng, 2007; Liou et al., 2011b; Liu and Gao, 2007; Nejati et al., 2009; Pakdil and Aydin, 2007; Sultan and Simpson, 2000; Tsai et al., 2011; Tsaur et al., 2002	Awasthi et al., 2011; Fernandes and Pacheco, 2010; Kuo, 2011; Nathanail, 2008; Sánchez et al., 2007; Yeh and Kuo, 2003; Yeh et al., 2000

[a] This is a representative, but not a comprehensive list of references; [b] Most of them try to develop a CSI or SQI; and [c] Based on disconfirmatory theory (Parasuraman et al., 1988)

Eq. 1 implies that all the attributes are equally important or have the same weight in SQ. Other authors have proposed weighting each attribute by a weight that would take the importance of each attribute in SQ into consideration. This would give a weighted SERVQUAL. Pakdil and Aydin (2007) used this method for analyzing airline SQ with loadings derived from factor analysis. Recently, Chou et al. (2011a) included fuzziness in SQ evaluation by using a fuzzy weighted SERVQUAL to evaluate airline SQ in Taiwan.

Although SERVQUAL represents the most widely adopted method for measuring SQ, the scale for capturing customer judgments has some disadvantages in obtaining an overall numerical measure of SQ. In fact, to calculate an index, analysts are forced to assign a numerical code to each level of judgment. In this way, equidistant numbers are assigned to each qualitative point of the scale, thus presuming that the distances between two consecutive levels of judgment expressed by the customers have the same size.

Another measure for SQ evaluation is provided by the Customer Satisfaction Index (CSI) (Hill et al., 2003). CSI represents a measure of SQ on the basis of attributes' importance rates and satisfaction rates (see Eq. 2).

$$\mathrm{CSI} = \sum_{k=1}^{N} [\bar{S}_k . W_k] \qquad (2)$$

where S_k is the mean of the satisfaction rates expressed by users on the service quality k attribute; and W_k (importance weight) is a weight of the k attribute, calculated on the basis of the importance rates expressed by users. Specifically, it is the ratio between the mean of the importance rates expressed by users on the k attribute and the sum of the average importance rates of all the service quality attributes.

CSI represents a good measure of overall satisfaction because it summarizes customer judgments on several service attributes in a single score. However, customer satisfaction rates can be very heterogeneous among users. These heterogeneities cannot be taken into account in the CSI calculation. To overcome this lack, importance weights and satisfaction rates can be corrected according to their dispersion. Eboli and Mazzulla (2009) introduced these adjustments for calculating a Heterogeneous Customer Satisfaction Index (HCSI) that they used to evaluate two suburban bus lines in Cosenza (Italy). HCSI was calculated by Eq. 3.

$$\mathrm{HCSI} = \sum_{k=1}^{N} [S_k^c . W_k^c] \quad \text{where} \quad S_k^c = \bar{S}_k . \frac{\frac{\bar{S}_k}{var(S_k)}}{\sum_{k=1}^{N} \frac{\bar{S}_k}{var(S_k)}} \cdot N \quad \text{and} \quad W_k^c = \frac{\frac{\bar{I}_k}{var(\bar{I}_k)}}{\sum_{k=1}^{N} \frac{\bar{I}_k}{var(\bar{I}_k)}} \qquad (3)$$

where S^c_k is the mean of the satisfaction rates expressed by users on the k attribute corrected according to the deviation of the rates from the average value; and W^c_k is the weight of the k attribute calculated on the basis of the importance rates expressed by users, corrected according to the dispersion of the rates from the average value.

HCSI introduces heterogeneity into user judgments. By introducing this adjustment, more significance is given to the attributes characterized by homogeneous user judgments, while less significance is given to the more heterogeneous attributes (Eboli and Mazzulla, 2009).

This group can also include Multicriteria Analysis when customers are asked for their degree of satisfaction with a specific criterion or attribute. Multicriteria Analysis has been widely used to deal with problems involving multiple criteria or attributes, as in the case of public transport quality of service analyses that involve multiple criteria of multilevel hierarchies and subjective assessments of decision alternatives. Frecuently, Multicriteria Analysis has been used combined with a fuzzy approach: Kuo et al. (2007) assessed SQ for interuban bus services in Taiwan and several authors (Chang and Yeh, 2002; Liou and Tzeng, 2007; Nejati et al. 2009; Tsaur et al., 2002) have analyzed SQ for airlines. In the model proposed by Chang and Yeh (2002) subjectivity could be taken into consideration in assessments in terms of attribute satisfaction and attribute importance. The model proposed by Liou and Tzeng (2007) takes into account that attributes are not ussualy independent. Tsaur et al. (2002) and Nejati et al. (2009) ranked airlines' SQ factors using a fuzzy TOPSIS approach. The approach is based on the idea that the selected alternative should have the shortest distance from the positive-ideal solution and the longest distance from the negative-ideal solution. Fuzzy TOPSIS extends TOPSIS to cases conducted in uncertain and fuzzy environment. The VIKOR method, which is based on an aggregate function representing "closeness to the ideal point" has also been adopted recently (Kuo and Liang, 2011; Liou et al., 2011b; Tsai et al., 2011). Opricovic and Tzeng (2004) compared VIKOR and TOPSIS and demonstrated that TOPSIS does not consider the relative importance of attributes.

2.1.2.2.- Aggregated models based only on performance

Cronin and Taylor (1992) criticised the measurement of SQ through gap model and proposed that perceptions only are better predictor of SQ. They developed a performance only measurement (SERVPERF) by illustrating that SQ is a form of consumer attitude and the performance only measure is an enhanced means of measuring SQ. Overall service quality is evaluated by performance perceptions only according to:

$$\mathbf{SQ} = \sum_{\mathbf{j=1}}^{\mathbf{k}} \mathbf{P_{ij}} \tag{4}$$

where k is the number of attributes; and P_{ij} is performance perception of stimulus i with respect to attribute j.

Such models have also been used in the field of public transportation. Sánchez et al. (2007) adapted the scale and proposed a weighted SERVPERF for assessment of the local bus service in Almería (Spain).

A multicriteria analysis based on SERVPERF was also used for analysing SQ in the public transport sector. Yeh et al. (2000) and Awasthi et al. (2011) used a fuzzy multicriteria analysis approach for evaluating urban transportation systems, posing questions on performance perception and importance. This approach was also used for airlines (Kuo, 2011) and airports evaluation (Fernandes and Pacheco, 2010; Yeh and Kuo, 2003 Fernandes and Pacheco, 2010; Yeh and Kuo, 2003). Nathanial (2008) evaluated the Hellenic railways using a multicriteria analysis based on objective and subjective data from several sources: statistical data, which are maintained by the railway operator; data which is collected through investigation conducted by trained personnel (mystery rider); and subjective data obtained through a CSS. In their opinion, it is impossible for a passenger to be able to provide a global performance grade of the itinerary based on a short experience, and therefore a combination of objective and subjective measurements is proposed (Transportation Research Board, 1999).

2.1.2.3.- Disaggregated models based on performance only

The disaggregated models most widely used to evaluate SQ are based on Quadrant Analysis (see Figure 1). Of these, the most widespread is Importance-Performance Analysis (IPA) (Martilla and James, 1977), which uses the attribute importance and the attribute performance as coordinates. This quadrant chart quantifies how important each attribute appears to be from a customer perspective (using the vertical axis) and shows the average customer rating for each characteristic (using the horizontal axis). Managers should focus on the position of each attribute in the four quadrant boundaries of the IPA matrix that show the relative urgency of improvement (Martilla and James, 1977). Higher ratings on the right side of the quadrant chart are better scores, and those on the left side are worse. The top-left quadrant identifies those attributes that appear to be most important but that are rated relatively low.

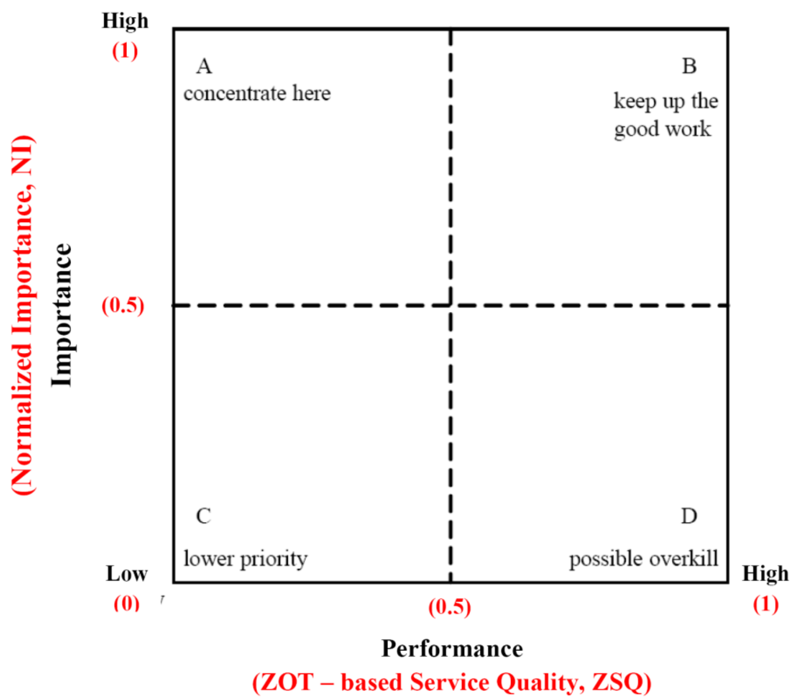


Figure 1. IPA matrix and NIZSQ matrix

This simple technique not only prescribes the prioritization of attributes for improvement, but also provides guidance for strategy formulation (Burns, 1986; Sampson and Showalter, 1999; Slack, 1994). However, the IPA matrix is a visualizing method and how to determine the priority of improving the SQ of the attributes remains unsolved. The precise ranking of the priority of improving attributes remains ambiguous and unidentified (Abalo et al., 2007).

IPA was applied to evaluate the BART (Bay Area Rapid Transit) rapid-transit system in San Francisco area (Weinstein, 2000) and the high speed railways in Taiwan and Korea (Chou et al., 2011b). Chen and Chang (2005) evaluated the airlines SQ in Taiwan using this method. Models of this type are widely used by transport company managers in the metropolitan transport sector (Christopher et al., 1999; Figler et al., 2011; Foote and Stuart, 1998) owing to their simplicity.

Stradling et al. (2007) introduced the user disgruntlement measure, derived by cross-tabulating performance against importance rating for each attribute[3]. They used a variation of the IPA to analyze different aspects of a particular service (e.g. user satisfaction with bus interchange), to compare across modes (e.g., user satisfaction with trips by car and bus), and within a mode across population sub-groups.

Eboli and Mazzulla (2011), following Nathanial (2008) and Tyrinopoulos and Aifadopoulou (2008), recently used a non-weighted disaggregated method, based on the use of both passenger perception and transit agency performance measures, to evaluate a suburban bus line. The method is based on each attribute having a subjective indicator (S) (calculated by the average of satisfaction rates expressed by a sample of users about the attribute) and an objective indicator (O) (obtained from performance indicators or, for the most qualitative attributes, calculated as the average of the scores assigned by operators or mystery riders to the parameters). Subsequently, through an optimization process, using the variance of S and O, a composite indicator (X) was obtained for each attribute. If the variance of the objective indicator is very low (close to 0) the X value coincides with the O indicator, by ignoring S indicator, and vice versa.

2.1.2.4.- Disaggregated performance-expectation models

Parasuraman et al. (1991b) proposed the concept of the zone of tolerance (ZOT) of expectations. They thought expectation could be divided into two levels of customer expectation: desired service (DS) and adecuate service (AS). ZOT is the difference between DS and AS (Parasuraman et al., 1991b), service superiority (SS)

[3] Disgruntled users for an attribute are those who consider that an attribute is highly important or very highly important, and at the same time considers its performance to be poor or very poor (Stradling et al., 2007).

is the difference between DS and percived service (PS), and service adequacy (SA) is the difference between PS and AS (Zeithaml et al., 1993) (See Figure 2). Following Parasuraman et al. (1991b), DS is the service the customer hopes to receive (it is a blend of what the customer believes "can be" and "should be") and AS level is that which the customer finds acceptable. If PS is lower than the ZOT, the consumer will have a negative evaluation of the service. When the PS is inside the ZOT, the consumer can be willing for buying, although he/she can also change to others service providers. And if the PS is higher than the DS, the service provider can delight consumers and improve their loyalty (Santos and Boote, 2003).

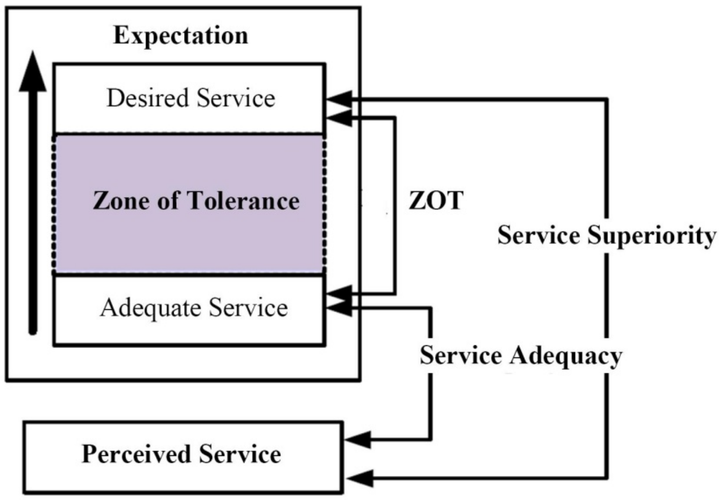


Figure 2. Zone of Tolerance of Zeithaml et al. (1993) (Source: Hu, 2010)

Hu and Jen (2006) define service quality in terms of the difference between perceived quality and tolerable quality (AS following Parasuraman et al. (1991b)) and apply it to evaluate the SQ of city buses in Taipei. Cavana et al (2007) use the ZOT for managing passenger rail service quality in New Zealand. Recently, Chang et al. (2012) introduced the fuzzy ZOT concept and applied it in the airline cargo business in Taiwan.

IPA has also been used changing performance by satisfaction (Mathisen and Solvoll, 2010; Wang et al., 2010). Recently, Tsai et al. (2011) combined AHP, VIKOR and IPA methods for considering airport passengers preferences (importance) and satisfaction simultaneously. The AHP was employed to measure the relative importance of each attribute;, the VIKOR method was used for computing the customer gaps of airport passenger service. And, finally, IPA was used for improving (reducing the gaps) attributes with higher importance.

Based on ZOT and IPA, Hu (2010) proposed the concept of zone of tolerance of expectation for evaluating SQ (ZSQ) and built an analytical framework for prioritizing attributes through a QA based on ZSQ and normalized importance (NIZSQ method). ZSQ is based on the concept of the 'performance ratio' in the customer satisfaction area (Vavra, 1997). The 'performance ratio' quantifies how much, from minimal to superior performance, an organization has progressed on a specific attribute. According to the same concept, ZSQ can show the 'service quality ratio'. Since DS, AS and PS can be seen as 'superior', 'minimum' and 'current', ZSQ can be expressed by the following equation:

$$\mathrm{ZSQ} = \frac{\mathrm{PS-AS}}{\mathrm{DS-AS}} = \frac{\mathrm{SA}}{\mathrm{ZOT}} \tag{5}$$

The meaning of SA divided by ZOT represents the performance ratio of SQ according to the customers' expectation. The smaller value of the service attribute's ZSQ means worse performance and should therefore have a higher priority to be improved.

After evaluating the ZSQ, managers need to consider the attribute's importance for judging the priority for improving attributes whose ZSQ values are between '0' and '1'. They only need to focus on values between '0' and '1' for two reasons (Hu, 2010):

- If ZSQ >1, PS is higher than DS and there is no need for improvement at the moment.

- If ZSQ < 0, the attribute must be improved immediately without any prioritizing analysis.

The Normalized Importance–ZSQ Analysis (NIZSQ) can be used for this purpose (Hu, 2010). NIZSQ method normalizes the importance data (NI) and replaces the x-axis in the IPA by ZSQ. Thus, NIZSQ analysis can be drawn as a two-dimensional diagram whose x-axis and y-axis have the same range (see Figure 2). Since the ranges of both NI and ZSQ are from '0' to '1', they can be divided into four quadrants by the mid-point at 0.5. The meanings of the four quadrants are the same as for the traditional IPA.

The top-right and bottom-left diagonal shows the ideal positions for attributes, which means that the performance of service quality is even with the importance (Slack, 1994). So, attributes on the left side of the diagonal need to be improved. The horizontal distance between attributes and the diagonal represents the improving space and the degree of urgency. The longer the distance is, the larger is the space to improve, and therefore the higher the priority to be improved. Thus, the horizontal distance (*d*) is used to calculate the prioritization. If the attributes have the same d value, they should be prioritized by their importance (Hu, 2010).

Hu (2010) used NIZSQ analysis to evaluate SQ of bus service in Taipei and compared the results with a traditional IPA. While IPA may lead managers to focus only on some items and ignore others, NIZSQ analysis reminds managers that they should keep those items in mind. Furthermore, NIZSQ analysis is not only a quadrant analysis, but also offers the improvement priority (d value) of each item based on the ZOT. Since prioritization is critical to managers' planning and they are usually unlikely to be able to focus on all items, *d* value can give them clear information regarding which items should be improved in priority and which items later.

2.1.2.5.- Other analyses

Finally, there are other studies in the literature that do not come under any of the methodological approaches indicated in Table 4.

A very high number of studies are conducted in terms of verifying hypotheses on SQ based on data supplied by CSS. Some of the studies use standard statistical methods (e.g. t-test, ANOVA, MANOVA, etc.) to confirm the hypotheses (Drea and Hanna, 2000; Ganesan-Lim et al., 2008; Oyewole, 2001; Pedersen et al., 2011). Most of them, however, use more advanced methods, such as Structural Equations Models (SEM) (Andreassen, 1995; Chen, 2008; Cheng et al., 2008; Chou and Kim, 2009; Chou et al., 2011b; Friman and Gärling, 2001; Friman et al., 2001; Jen et al., 2011; Joewono and Kubota, 2007a; Joewono and Kubota, 2007c; Kim and Lee, 2011; Lai and Chen, 2011; Minser and Webb, 2010; Park et al., 2006; Saha and Theingi, 2009; Stuart et al., 2000; Tripp and Drea, 2002; Wen et al., 2005; Yang et al., 2012) or Path Analysis (Forgas et al., 2010; Jen and Hu, 2003; Joewono and Kubota, 2007b; Lin et al., 2008; Park et al, 2004; Ringle et al., 2011) for verification purposes[4].

Other papers study the differences in service quality perceived by different groups of individuals, services or companies, or before/after carrying out an action. An important number of authors (Aksoy et al., 2003; Drea and Hanna, 2000; Ganesan-Lim et al., 2008; Gilbert and Wong, 2003; Glascock, 1997; Kim et al., 2011; Koushki et al., 2003; Murray et al., 2010; Oyewole, 2001; Paquette et al., 2012; Park, 2007; Pedersen et al., 2011; Ritchie et al., 1980; Surovitskikh and Lubbe, 2008; Wen et al., 2008; Tyrinopoulos and Antoniou, 2008) compare SQ through different categories of users or population groups, using the standard statistical methods pointed out in the previous paragraph. This type of analysis has been also conducted using more advanced methods, including ordered choice models (Dell'Olio et al., 2010; Hensher et al., 2010; Huse and Evangelho, 2007), SEM (Andreassen, 1995; Friman et al., 2001), or Path Analysis (Ringle et al., 2011).

Most authors compare SQ in different services and companies to each other, using standard statistical methods (Aksoy et al., 2003; Drea and Hanna, 2000; Kim et al., 2011; López-Bonilla and López-Bonilla, 2008; Ostrowski et al., 1993; Park, 2007; Surovitskikh and Lubbe, 2008; Tyrinopoulos and Antoniou, 2008), although

[4] SEM and Path Analysis are described in Section 2.1.3.

comparisons have also been made using SEM (Chou and Kim, 2009; Chou et al., 2011b), Path Analysis (Forgas et al., 2010) and ordered choice models (Tyrinopoulos and Antoniou, 2008).

Finally, some studies analyzed SQ before and after carrying out an action (Foote et al., 2001; Friman, 2004; Pedersen et al., 2011) using standard statistical methods (e.g. t-test, ANOVA, MANOVA, etc.).

2.1.3. Approaches to estimating the relative importance of each service quality attribute

Section 2.1.2 shows that most approaches use the importance of each attribute. Public transport companies want to know not only how their customers rate them on detailed service attributes (attribute-performance ratings), but also the relative importance of these attributes (attribute-importance measures) to their customers. The most widely used approach (see Table 3) is asking customers to rate each attribute on an importance scale (Stated Importance), although methods that derive attribute importance by statistically testing the strength of the relationship of individual attributes with overall satisfaction (Derived Importance) are also widely used.

2.1.3.1. Stated importance

Stated importance is the more intuitive and simpler of the two methods: besides being asked to assess each attribute, users are asked to indicate the importance the attributes have for them.

However, this approach has several disadvantages (Weinstein, 2000):

- it requires a significant increase in the length of the survey instrument. This can depress the overall response rate and accuracy of the survey,

- it can sometimes yield insufficient differentiation among mean importance ratings, with customers rating nearly all of the measures near the top of the scale; or

- attributes may be rated as important even though they in fact have little influence on overall satisfaction.

In some cases (Liou and Tzeng, 2007; Tsai et al., 2011; Tsaur et al., 2002) more sophisticated processes are used, such as the Analytic Hierarchy Process (AHP) of Saaty (1994). In other cases, the questions were posed not only to users but also to transport companies and government (e.g. Yeh et al., 2000).

2.1.3.2. Derived importance

It is common practice to include in CSS both questions about a customer's overall satisfaction with the service and detailed questions about specific characteristics of the service. The information gathered can be used in several statistical methods (e.g. bivariate correlations, multiple-regression analysis, SEM, etc.) for deriving the attributes weight or importance from CSS.

2.1.3.2.1. Factor analysis

Factor analysis is a set of multivariate statistical techniques whose primary goal is to investigate whether a number of variables of interest are linearly related to a smaller number of unobservable factors. Factor analysis is related to principal components analysis (e.g., both rely on the correlation matrix), but the two are not identical. In factor analysis, the researcher makes the assumption that an underlying causal model exists, whereas principal components analysis is simply a data reduction technique.

Factor analysis provides a better understanding of how customers perceive various service attributes by showing which attributes tend to be thought of similarly. This technique is normally used as a preliminary step for other methods, such as multiple linear regression analysis (Kim and Lee, 2011; Weinstein, 2000), discriminant analysis (Aksoy et al., 2003) or SEM (Eboli and Mazzulla, 2007). The factors provide a more manageable number of variables with which to carry the analysis to the next level.

2.1.3.2.2. Bivariate correlations

Bivariate correlations can be used as a tool for ranking the relative importance of each attribute (Figler et al., 2011; Weinstein, 2000). Those authors have calculated bivariate correlation coefficients between each attribute's rating and the overall satisfaction rating to estimate the importance of each service characteristic. The main disadvantage of this method is that it disregards the correlation among attributes, so it is important not to interpret the coefficients too literally owing to the extensive colinearity among them.

2.1.3.2.3. Regression analysis

The purpose of regression analysis is to assess the relative importance of each factor and to test the overall explanatory power of the battery of factors as a whole. In the regression model, the factors serve as the independent variables, whereas overall satisfaction, or SQ, serves as the dependent variable. Regression analysis results in a best-fitting model in the form of an equation that expresses the dependent variable as a combination of the independent variables. Several models of regression have been proposed to study satisfaction or SQ (see Table 3). Papers based on multiple linear regression models (Kim and Lee, 2011; Weinstein, 2000) do not take the categorical nature of the dependent variable into consideration and are infrequently used in the literature. The most widely used methods are the ones that take into account that the dependent variable is categorical.

Aksoy et al. (2003) propose using Discriminant Analysis to identify key service dimensions for predicting satisfaction in airlines. Discriminant Analysis undertakes the same task as multiple linear regressions by predicting an outcome, but considering that the dependent variable is categorical. Logistic regression and probit regression are similar to Discriminant Analysis, as they also explain a categorical variable. However, these other methods are preferable in applications where it is not reasonable to assume that the independent variables are normally distributed, which is a fundamental assumption of the Discriminant Analysis method. Several authors propose Ordered Logit (Tyrinopoulos and Aifadopoulou, 2008; Tyrinopoulos and Antoniou, 2008) and Ordered Probit (Dell'Olio et al., 2010; Huse and Evangelho, 2007) models to study the relationship between overall satisfaction and each of the attributes or factors under consideration. Ordered Logit or Ordered Probit models are extension of the logistic or probit regression models, allowing for more than two (ordered) response categories, which is the situation encountered with the CSS. In Ordered Probit models the unobserved terms are supposed to be distributed standard normal instead of logistic, which is the hypothesis in Ordered Logit. Recently, Castillo and Benitez (2012) applied a Generalised Linear model for modeling the global satisfaction of a bus public service and Hensher et al. (2010) proposed a Generalized Ordered Logit model that accounts for preference heterogeneity through random parameters.

2.1.3.2.4. Structural equation models

In recent years, the Structural Equations Models (SEM) have been widely used. SEM is a multivariate technique combining regression, factor analysis, and analysis of variance to estimate interrelated dependence relationships simultaneously. This approach allows the modelling of a phenomenon by considering both the unobserved "latent" constructs and the observed indicators that describe the phenomenon.

SEMs are made up of two elements: the first describes the relationship between endogenous and exogenous latent variables, and permits the evaluation of both direction and strength of the causal effects among these variables (structural model); the second component describes the relationship between latent and observed variables (measurement model).

The basic equation of the structural model is defined as (Bollen, 1989):

$$\eta = B\eta + \Gamma\xi + \zeta \tag{6}$$

in which η is a *mx1* vector of the latent endogenous variables, ξ is a *nx1* vector of the latent exogenous variables, B is a *mxm* matrix of the coefficients associated with the latent endogenous variables, Γ is a *mxn* matrix of the coefficients associated with the latent exogenous variables and ζ is a *mx1* vector of error terms associated with the endogenous variables.

The basic equations of the measurement model are the following:

$$x = \Lambda_x \xi + \delta \tag{7}$$

$$y = \Lambda_y \eta + \varepsilon \tag{8}$$

in which x and δ are column q-vectors related to the observed exogenous variables and errors, respectively; Λ_x is a *qxn* structural coefficient matrix for the effects of the latent exogenous variables on the observed variables, y and ε are column p-vectors related to the observed endogenous variables and errors, respectively, and Λ_y is a *pxm* structural coefficient matrix for the effects of the latent endogenous variables on the observed ones.

The structural equation system is estimated by using different methods: maximum likelihood, weighted and un-weighted least squares, generalized least squares, and so on. All of them are based on the covariance analysis method, in which the difference between the sample covariance and the model implied covariance matrices is minimized. The maximum likelihood method is the most popular, however selecting an appropriate SEM estimation method depends on different assumptions about the probability distribution, the scale properties of the variables, the complexity of the SEM, and the sample size For a more detailed discussion on SEM and estimation methods see Joreskog (1973), Bollen (1989), Golob (2003) and Washington et al. (2003).

SEM were generalized by Joreskog (1973) and Wiley (1973). Many applications have been proposed in several fields of research: psychology, social science, natural science, economics, statistics, etc. SEM has also been adopted for describing customer satisfaction in several public transport services: metropolitan public transportation (Andreassen, 1995; Eboli and Mazzulla, 2007; Friman and Gärling, 2001; Friman et al., 2001; Joewono and Kubota, 2007a; 2007c; Karlaftis et al., 2001; Lai and Chen, 2011; Minser and Webb, 2010; Nurul-Habib et al., 2011; Stuart et al., 2000); interurban bus services (Jen et al., 2011; Wen et al., 2005); rail transportation (Chou and Kim, 2009; Chou et al., 2011b; Tripp and Drea, 2002) and airlines (Chen, 2008; Cheng et al., 2008; Kim and Lee, 2011; Park et al., 2006; Saha and Theingi, 2009; Yang et al., 2012).

Path analysis can be viewed as a special case of SEM: one in which only single indicators is employed for each of the variables in the causal model. That is, path analysis is SEM with a structural model, but no measurement model. Several authors (Forgas et al., 2010; Park et al, 2004; Ringle et al., 2011) used this method for modelling airlines SQ. Jen and Hu (2003) and Lin et al. (2008) have used path analysis for evaluating bus services in Taiwan. Finally, Joewono and Kubota (2007b) used it for analysing user perceptions of paratransit in Indonesia.

2.1.4. Summary and discussion

The bibliography (see Table 1) indicates that the first studies of SQ in the PT sector emerged in the air transport sector and in urban and metropolitan public transport in the late 20th century. However, such studies have increased considerably at the start of the 21st century, particularly in the field of airlines and urban and metropolitan transport. The two sectors pose different problems but share the same goal: to increase the number of passengers. In the case of air transport, the deregulation and opening-up-the-sky policies of the airline industry have put pressure on airlines and airports worldwide to become more competitive. In the urban and metropolitan public transport, companies and governments are highly interested in enhancing the quality of public transport in order to discourage the use of private vehicles. There have been fewer documented studies on ground interurban public transport (by bus or by train). It is to be hoped that SQ concerns in this sector will grow when public transport services are truly deregulated, which is one of the European goals.

The methodological approaches most widely used by practitioners, transport operators and governments are the ones based on CSS that use a quadrant analysis, such as the IPA and its variants (e.g. IPA with satisfaction, NIZSQ, etc.) (Christopher et al., 1999; Department for Transport of England, 2005; Figler et al., 2011; Foote and Stuart, 1998; Weinstein, 2000; Transport for London, 2006). Such methods are encompassed in what are known as disaggregated models. They help managers to set priorities for service improvements among a long list of service attributes. On the contrary, the approaches preferred by researchers and academics have sought to arrive at a global indicator (SQI or CSI) that could be used to compare different services and their development over time. The most widely used models in this case have been based on disconfirmation theory (Parasuraman et al., 1988). In recent years, however, both approaches (aggregated and disaggregated) are being used to complement each other.

Given the subjective, qualitative and imprecise nature inherent to SQ evaluation data, a growing number of studies, particularly in the field of air transport, are using fuzzy set theory as an effective way for formulating this kind of problems (Awasthi et al., 2011; Chang and Yeh, 2002; Chou et al., 2011a; Fernandes and Pacheco, 2010; Kuo and Liang, 2012; Kuo, 2011; Nejati et al., 2009; Yeh and Kuo, 2003; Yeh et al. 2000).

One emerging trend in recent years proposes the combined use of subjective information obtained from users (through CSS) and objective data on service performance supplied by transport companies and governments (Eboli and Mazzulla, 2011; 2012; Nathanail, 2008; Tyrinopoulos and Aifadopoulou, 2008; Yeh et al. 2000). This new approach is based on the consideration that passengers' perceptions alone can lead to many biases, especially when users' judgements are too heterogeneous. In addition, subjective measures are based on users' opinions, while non-users' perceptions are not considered. On the other hand, a specific objective indicator could be not appropriate for evaluating a service aspect, or could fail to fully describe a service aspect characterized by various factors (Eboli and Mazzulla, 2011).

Analysing the heterogeneity of user perceptions has focused attention from the first studies on SQ in the public transport sector (e.g. Glascock, 1997; Ritchie et al., 1980) in which the difference in the SQ perceived by different groups of individuals that had been previously segmented (business/vacation, riders/non-riders, etc.) were compared using standard statistical methods. In recent years, further studies on the heterogeneity have used more sophisticated models, including generalised ordered choice models (Hensher et al., 2010), SEM (Andreassen, 1995; Friman et al., 2001) or Path Analysis (Ringle et al., 2011) after stratifying samples in terms of socioeconomic, demographic and population habits data. More recently, several works (Huse and Evangelho, 2007; Wen and Lai, 2010; Wen et al., 2008) have proposed new approaches that use several methods to identify clusters. Thus, the population can be segmented and the heterogeneous preferences can be studied. This new method can be used to study specific population segments whose behaviour could hardly be identified by conventional stratification based on socioeconomic and/or demographic factors.

Heterogeneous Customer Satisfaction Index proposed by Eboli and Mazzulla (2009) is also interesting with regards to take into account the heterogeneity of user perceptions. It can be used to give more significance to the SQ attributes characterized by homogeneous user judgement, while less significance is given to the more heterogeneous attributes.

Moreover, as indicated previously, discrete choice models based on stated preference surveys are becoming more widespread in the analyses of SQ in the public transport sector. Several studies (Eboli and Mazzulla, 2008; Hensher and Prioni, 2002; Hensher et al., 2003; Gatta and Marcucci, 2007) assume that the overall level of passenger satisfaction is best measured by how an individual evaluates the total package of services offered. Appropriate weights attached to each service dimension will reveal the strength of positive and negative sources of overall satisfaction. The weights are estimated using several models based on stated preference surveys, such as: multinomial logit, hierarchical or nested logit and mixed logit models. Specifically, mixed logit models have been used in recent years because they can consider the heterogeneity of perceptions.

The importance of SQ attributes on global customer satisfaction can be evaluated by the estimation of coefficients of the models. Gatta and Marcucci (2007) point out that these methods overcome some critical factors pertaining to methods based on CSS, such as conceptual grounds, psychometric problems and troubles with Likert scales. The latter, in particular, have a well-documented tendency for respondents to choose central response options rather than extreme ones. Other factors include the impact of the number of scale points used; the influence of the format and verbal labelling of the points; and the transformation from ordinal data to cardinal data.

The drawbacks of the methods that determine the relative importance of each SQ attribute based on stated importance are many: significant increase in the length of the survey; insufficient differentiation among mean importance ratings, with customer rating nearly all of measures near the top of the scale; and attributes that are rated as important even though they have little influence. Nonetheless, there has been no increase in the methods used to derive importance by statistically testing the strength of the relationship of individual attributes with overall satisfaction. This may be largely due to the fact that although statistically inferred methods can overcome the shortcomings of stated importance ratings, most of them carry the assumptions of relatively normal data, linear relationships between independent and dependent variables, and the relatively low multi-colinearity between independent variables and, in customer satisfaction research, these assumptions are almost always violated (Garver, 2003). Lately, new methods are being proposed that can overcome these weaknesses, such as the AHP (Liou and Tzeng, 2007; Tsai et al., 2011; Tsaur et al., 2002) and others based on decision trees (De Oña et al., 2012).

The methods based on decision trees have the advantage of not needing assumptions or pre-defined underlying relationships between dependent and independent variables, and therefore they could be used to study SQ in the public transport sector.

2.2. Decision Trees

2.2.1. What is a Decision Tree?

Decision trees (DTs) is a data mining technique used for the classification and prediction of a target variable. This technique has the ability to discover useful patterns in great amounts of data that allows us to make predictions on new data. Nowadays, the quantity of information that the current society can use increases day a day, generating huge data sets, which real worth resides in the information that could be extracted from them. DTs are popular due to their ability to handle large amounts of data and also to their ability for extracting the hidden knowledge from these huge data sets. Their success emerges also due to their simplicity and transparency; in fact, they are usually presented graphically as hierarchical structures, which make them easy to interpret.

There are two types of DTs according to the nature of the target variable. When the value of the target variable is discrete, a classification tree is developed and the outcome to be predicted is a discrete class, whereas when the value of the target variable is continuous, a regression tree is generated and a numeric quantity is predicted.

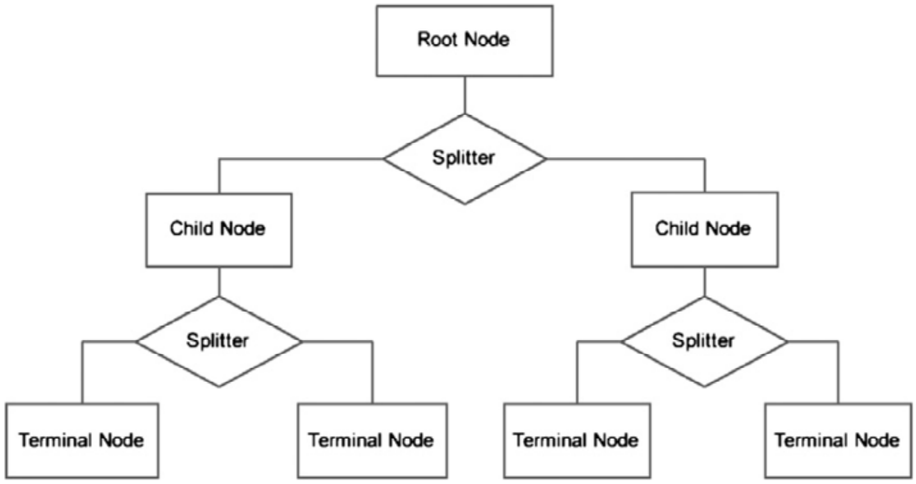


Figure 3. General structure of a decision tree (Source: Kashany and Mohaymany, 2011)

The structure of a DT is represented by nodes and branches (see Figure 3). There are three types of nodes inside a DT:

- *Root Node.* It is the node located at the top of the tree, which has no edges enter, and all the data are concentrated on it.

- *Child Nodes,* which are internal nodes that involve testing a particular attribute.

- *Terminal Nodes or Leaf Nodes* which have no branches and assign a classification or prediction to all the instances which reach the leaf.

The branches represent one of the states or values of the attribute that is used as Splitter. To classify an unknown instance, it is routed down the tree according to the values of the attributes tested in successive nodes, and when a leaf is reached the instance is classified or predicted according to the class or value assigned to the leaf (Witten and Frank, 2005).

The building process of a DT follows a descending strategy (top-down), which is based on the principle “divide-and-conquer”. Starting from a set of data (which constitute the root node), and using specific splitting criteria, it begins selecting an explanatory attribute to place at the root node, and create two or more branches for the different values of the attribute, splitting up the full data set into smaller subsets.

When the explanatory variable used as splitter is nominal, the number of branches created depends on the algorithm used. It can produce multiway-splits or binary splits. In the first case, usually the number of branches is equal to the number of possible categories of the splitter and the independent variables are used only once as splitters further down the tree. For binary splits, the number of branches are always two, joining the categories of the independent variables into two groups looking for the optimal partition. In binary splits the variables might be retested more than once in the path from the root node to a leaf. For numeric variables, the condition refers to determining whether the value of the attribute is greater or less than a predetermined threshold. Usually, numeric explanatory variables produce binary branches independently of the algorithm used, although also multiway-splits can be performed. They are often retested in the path from the root node to a leaf, according to different threshold values.

The splitting process can be repeated recursively for each branch, using only those cases that actually reach the branch. The objective is to obtain more homogeneous subsets (in terms of the target variable) with each split carried out. Then, each subset is split recursively until all of them are pure (when the cases in each subset are all of the same class) or a stopping criteria has been satisfied.

Subsequently the tree stops growing, terminal nodes are created. Each terminal node is assigned to one class representing the most appropriate target category, or it is predicted a value of the target variable. Alternatively, the terminal node may hold a probability vector indicating the probability of the target variable having a certain value.

However, following this process, a saturated tree is obtained. The saturated tree provides the best fit for the data set which it is constructed from, but overfits the information contained within this data set. This overfitting does not help in accurately classifying another data set. In order to solve this problem, the tree can be pruned. The pruning process involves finding, from the complex tree generated, a simpler model that also fits the data without overfitting them. Generally, it causes that the accuracy on the data set used for building the tree decreases, but it will increase the accuracy on other new data set. Then, at each node of the tree, using a specific pruning criteria, it has to be decided if prune the tree or leave it as it is, unpruned.

According to Witten and Frank (2005), there are two different operations that can be considered for pruning, the subtree replacement and the subtree raising. The subtree replacement is based on replace some subtrees of the whole tree by single leaves, while the subtree raising works bringing up a below subtree to replace an upper subtree. This second pruning operation is more complex and it is not clear that it is necessarily always worthwhile.

On the other hand, establishing strict stopping criteria can also prevent to create saturated trees that overfit the data (sometimes this is called a prepruning process). However, the pruning process is preferred to do not leave growing the subtrees, because sometimes the attributes individually seem to have nothing to contribute, but when two or three of them are combined, they can be powerful predictors. If the tree model does not have a complete growth, this significant contribution may be not discovered.

Therefore, once the tree has been built and pruned, new instances can be classified following the path from the root of the tree down to a leaf, according to the outcomes of the tests along the path. Moreover, the decision tree built can easily be converted into decision rules. From each teminal node in the model, a Decision Rule is generated. Decision rules are conditional statements created by following a path from the root node to a terminal node. It provides a prediction of the target variable (a class or a value, depending on the nature of this variable) when a set of conditions are complied. They have the ability to explain the reasons for a decision, not interpreted as a direct causation, but as associations between sets of variables. They have been used to discover patterns such us two or more attributes that are often together. Then, the information extracted by the decision rules could provide useful and interesting insights of service quality for transport managers and providers, and permit them to formulate new policies and strategies based on this information. Later, DTs allow detecting patterns in a new sample, or simply to gain a better understanding of the phenomenon being analyzed.

In addition, the importance of the independent variables over the target variable can be derived from this methodology. In an aggregated analysis (as it has been pointed out before), knowing the weight or importance of each attribute in terms of global quality is essential for constructing an overall service quality index.

2.2.2. Methods for building Decision Trees

There are many different algorithms that can be used to generate DTs. The main difference between them lies in the splitting criteria used for the tree growth as well as the algorithm followed in the pruning phase. Some of the most popular algorithms used in different research fields are CHAID (Kass, 1980), CART (Breiman et al. 1984), ID3 (Quinlan, 1986), C4.5 (Quinlan, 1993), C5.0 (Quinlan, 1997), and so on. In this section, we do not try to determine which is the best algorithm, issue for which it does not exist consensus among researchers. Only a brief description of various of them is developed below.

The Chisquare–Automatic–Interaction–Detection (CHAID) algorithm was developed by Kass (1980). This algorithm performs the tree growth by non-binary splits, and the non-significant categories of the variables used as predictors are grouped with the significant categories in a merging process. The merging process and the splitting criteria applied in this algorithm depend on the nature of the target variable. When the target variable is continuous, a F test is developed, when it is nominal a Pearson chi-squared test, and when it is ordinal, a likelihood-ratio test is used. These criteria are used to determine for each predictor the categories that are not significant. It is found the pair of categories that is least significantly different, most similar, which is the one that achieves the highest p-value in the statistical test applied, with respect to the dependent variable. If the respective p-value for a given pair of categories is not statistically significant (is higher than a establish α value), it will merge the respective predictor categories and repeat this step. Therefore, the predictor variable choosed as splitter will be the one that have the smallest adjusted p-value (adjusted p-value computed by applying Bonferroni), that is the one that yields the most significant split. CHAID only accepts categorical predictors and when continuos variables are used as independent variables, they have to be recoded into categorical ones. Moreover, it does not perform the pruning process in the tree and the data used to build the tree can be overfitted. However, one of its advantages is that missing values can be handled treating them as a single valid category.

The Classification and Regression Trees (CART) algorithm was developed by Breiman et al. (1984). Depending on the nature of the dependent variable, a classification tree is created (discrete variable), or a regression tree is built (continuous variable). It yields binary trees, splitting the branches just in two ways. It uses the Gini Index as the splitting criteria, and the obtained tree is pruned by a Cost–Complexity algorithm. The Gini[5] index measures the degree of impurity in the nodes, evaluating the reduction of impurity between the parent node and the child nodes created with the possible splitters. One of the main advantages of CART methodology is that it can handle numeric and continuous attributes. However, due to the binary splits, its main disadvantage is that when the independent variables are composed by multiple categories (more than two), the influence of a specific category on the target variable is difficult to be analyzed, because the independent variables' categories are merged into two groups for achieving a binary partition. Only if this variable is retested along the path top-down the tree, the influence of a single category coul be identified.

The Iterative Dichotomiser 3 (ID3) algorithm is considered a very simple algorithm. It was developed by Quinlan (1986), who established the Information Gain as the splitting criteria. This algorithm also generates non-binary trees as the CHAID algorithm. ID3 developes the tree growing pocess by splitting the sample that reaches each node by a number of branches equal to the number of categories of the attribute used as splitter. The Information Gain is based on the entropy

[5] The Gini index is explained in Chapter 4

function, and represents the amount of information that will be necessary for classifying a new instance until a specific node. Then, the Information Gain can be defined as follows (Thill and Wheeler, 2000):

$$\mathrm{Gain_x(T) = Info(T) - Info_x(T)} \tag{9}$$

Where, *Gain*$_x$*(T)* is the Information Gain in a node *T* produced by a variable *X*, *Info (T)* is the entropy or information needed to identify the classification of the target variable at the node *T*, and *Info*$_x$*(T)* is the entropy after partitioning the node *T* into *n* classes according to the value(s) of a given attribute *X*. Each partition at each node in the decision tree is tested by the splitting criteria according to the different independent attributes and the cases of the sample that have passed down to that node. So, *Info (T)* and *Info*$_x$*(T)* are defined as follows:

$$\mathrm{Info(T)} = -\sum_{j=1}^{k} \frac{C_j}{n_T} \log_2\left(\frac{C_j}{n_T}\right) \tag{10}$$

$$\mathrm{Info_x(T)} = \sum_{i=1}^{n} \frac{n_{T_i}}{n_T} * \mathrm{Info}\ (T_i) \tag{11}$$

Where k is the number of classes of the target variable, C_j is the number of cases from the class j, n_T is the total number of cases in the node T, and n_{Ti} is the number of cases in the node T_i (the node generated by one of the n states of the variable X). Then, the principle followed by the Information Gain is to select the attribute that best minimizes the information necessary to correctly classify the observations, is that the Information Gain is maximal. A lower value of the entropy, less uncertainty and more useful will be the attribute for the classification. However, the Information Gain criterion tends to favor tests on attributes that have a large number of values. For instance, if a node have N cases, and the variable used for splitting have N single categories, *Info*$_x$*(T) =0* and the information gain is maximal (Thill and Wheeler, 2000). In addition to this disadvantage, this ID3 algorithm does not apply any pruning procedure to the tree (as the algorithm CHAID) and it does not handle numeric attributes or missing values.

Quinlan (1993) developed the C4.5 algorithm, as the successor of the ID3 method. This new algorithm introduced the Gain Ratio as splitting criteria and performed the pruning process. The Gain Ratio was also based on the entropy measure, but with an improved form. The Information Gain measure was biased in selecting the attributes for splitting the tree, due that the attributes with a higher number of categories were usually identified as the best splitters, rejecting the other ones. The Gain Ratio overcomes this problem using a normalized form of the Information Gain. The Gain Ratio is defined as (Thill and Wheeler, 2000):

$$\mathrm{GainRatio_x(T)} = \frac{\mathrm{Gain_x(T)}}{\mathrm{SplitInfo_x(T)}} \tag{12}$$

Where *SplitInfo*$_x$*(T)* is defined as follows:

$$\text{SplitInfo}_x(T) = -\sum_{i=1}^{n} \frac{n_{T_i}}{n_T} * \log_2\left(\frac{n_{T_i}}{n_T}\right) \tag{13}$$

Thus, at each node, the attribute that achieves the greatest value of the Gain Ratio is selected as the splitter. In this algorithm the pruning process is performed, and it usually is developed by a Error-based algorithm or a Pessimistic algorithm. The Pessimistic algorithm is prefered over the Error-based algorithm because it is faster without affecting its performance (Esposito et al., 1997). C4.5 algorithm also generates the same number of branches as the number of categories of the splitter when the independent variable is nominal, however, when the splitter is numerical, the number of branches in each split is limited to a binary partition as happens in the CART methodology. Then, its main advantages over the ID3 algorithm are: it performs the pruning process avoiding the overfitting of the data, it can handle numeric attributes and it also handles missing values. However, unlike CART or CHAID algorithm, it can not build decision trees for numerical target variables. Moreover, in general, the decision trees generated by this algorithm are bigger (have more branches) that the ones obtained by the CART algorithm.

The more recent version of this method is the C5.0 algorithm (Quinlan, 1997). It is the successor of the C4.5 algorithm and has some negligible advantages over this one. The main novelty of this method is that a boosting algorithm can be implemented, which improves the classification accuracy of the classifier. It estimates multiple iterative classifications, assigning, at each iteration, a weight to each observation. The observations that were misclassified in the previous iteration obtaine the heavier weight in order to force the classification algorithm to concentrate on those observations. Therefore, a new classification tree is created in each iteration, with the aim of correct the errors in the previous iteration. The other improvements of this algorithm are: the speed and accuracy are little higher, the computer memory usage is minor, it deals with continuous independent variables without branch limitations and it has the branch-merging option for nominal variables.

It should be stressed that each method has its advantages and disadvantages, and reveals different information. According to the algortihm used, they create binary or multiway trees, they use different splitting criteria (e.g. Gini index, Gain ratio, etc), the pruning phase is implemented or not, some of them can handle numeric attributes, others permit using missing values, and so on. Moreover, many researchers have pointed out that, in most cases, the choice of the splitting criteria will not make much difference on the tree performance. Then, there is no consensus on the best algorithm to be used. Researchers should focus on using the algorithm that best fits the characteristics of the data and the phenomenon to be analyzed.

However, in spite of this lack of consensus, CART methodology has been the most widely employed DT algorithm in many different studies, maybe due to its ability for treating with continuous target variables (developing regression trees), or due to the well reputation of the Gini index as the criterion for splitting the tree. It is not known the reason, but in the last years it has suffered a great spreading through different research fields, such as business administration, agriculture, industry, engineering and so on.

2.2.3. Applications of Decision Trees in transportation

Over the years, DTs have become a popular tool in many different research fields. Particularly, throughout the transport engineering field, this methodology has also achieved a fast spreading in the last years. The real boost in the usage of DTs in transport engineering began in the year 2000, when various researchers started to applying DTs in their investigations. However, to our knowledge, the first application of DT in transport engineering was in 1997, with the purpose of traffic forecasting and vehicle emissions analysis. The applications varied from those in transportation choice behaviour, road safety and so on. In the next sections a brief summary of these studies is presented.

2.2.3.1. Applications in transportation choice behavior

In the last decade, DTs have started to be used for predicting choice behavior in transportation. Spatial and travel choice behavior remains one of the most interesting and important areas of research in transportation planning, where discrete choice models have been widely applied. However, choice behavior can be well represented by DTs given that it can be considered as sequential examinations of attributes.

Thill and Wheeler (2000) used the C4.5 algorithm for modeling the choice among discrete travel destinations within the Minneapolis–St. Paul metropolitan area. They used the gain ratio as splitter criterion, and a pessimistic algorithm for pruning. They preferred the pessimistic pruning algorithm to the error-based algorithm, usually applied for the C4.5 model, because it was faster without affecting its performance. Moreover, they stated that the pessimistic pruning could ensure that the expected confidence levels in the training data were more similar to the actual confidence levels from unseen data, because if actual confidence levels are much worse than those predicted, it suggests that a better decision tree can be discovered.

For building the DT, they used 19 independent variables related to the separation between the trip origin (home) and the potential destinations (Time and Distance), the characteristics of potential destinations (population, employment, the form of urbanization, etc.), and traveler attributes (age, gender, income, etc.). The dependent variable was the shopping destination choice. Results showed that the trained model performed satisfactorily and achieved better results than other conventional discrete choice models applied on the same data. Furthermore, authors pointed out that the algorithm used in their research had the ability to uncover the structure of large and complex spatial databases and to build this knowledge into operational transportation decision support systems.

Analyzing the choice of trip chaining as a part of a multistop trip, was the purpose of Arentze et al. (2000) work. These authors consider that when a decision maker realizes an out-of-home activity, he/she either decides to conduct the activity on a separate home-based trip or as a multistop trip. They used three different algorithms for generating DTs: C4.5, CART and CHAID. The results were compared and showed better performances in C4.5 algorithm, followed by CHAID and in the last position was CART. However, they pointed out that C4.5 and CART were able to optimize the discretization of continuous attributes simultaneously, while CHAID algorithm could not.

Based on Revealed Preference surveys and Stated Preference surveys, a DT for predicting drivers' route choice behavior was modeled by Yamamoto et al. (2002). They affirmed that using DTs were better than using artificial neural networks (another data mining technique that had become very popular in the last few years for analyzing travel choice), because they facilitated the determination of the relationships between the explanatory variables and the choice. In their research work two models (with the C4.5 algorithm) were generated using data from two different cities in Japan and, in addition, two binary logit models were estimated in order to compare the results. The target variable for the models was defined with two alternative routes. The independent variables were the aspects related with drivers' expected travel times for each alternative route (minimum, maximum, and average travel times), the sociodemographic variables and the travel characteristics. The comparison of the precision ratios among the two methods (DT and binary logit model) showed better performance for the DTs in both study cases.

Lee et al. (2010) applied a hybrid methodology of DTs and an artificial neural network for analyzing the factors affecting car drivers' alternative route choice when variable-message signs are deployed. Data from a stated preference survey developed in Wisconsin (United States) was used. Authors also pointed out the important limitations of logistic regression models and artificial neural networks, widely used in this research field. Logistic regression models present difficulties in interpret the coefficients, and also in assuring that the model includes all relevant variables and excludes all irrelevant variables. The interpretation of artificial neural networks results and understanding of the behavioral characteristics of drivers, is even more difficult for artificial neural networks than for logistic regression models. The authors applied a hybrid tree model developed by Chan and Loh (2004), called LOTUS (Logistic Regression Trees with Unbiased Selection).

This method develops a tree model fitting different logistic regression models for each partition in the recursively splitting of the data. Comparison between the tree model and the neural network was carried out to validate and assess the performance of both methods. The findings indicated that the accuracy for the artificial neural network was slightly higher than for the tree model. However, the LOTUS model was more effective in analyzing the driver behavior data, with more interpretable results, and also had a reasonable prediction accuracy.

In order to have a comprehensive and clear description of the transport mode choice in the context of activity scheduling, Wets et al. (2000) applied DTs and a logit model to represent this behavior. They used two different algorithms to build the DTs (C4.5 and CHAID) and the results were compared with the logit model. The mode choice was categorized into three classes: Slow mode (including bike and walk), Pass (including car passenger and public transport) and carD (including only the category of car driver), and 41 independent variables related with the household/person level, the activity pattern level, the tour level, and the activity/trip level, were used to predict the choice. They discovered that the ratio of correctly predicted cases of a holdout sample was almost identical for the three methods, but the tree models had a potential advantage with respect to robustness. Whereas logit models assume a predefined form of the utility function (i.e., additive), tree models are theory-free, and their structure is derived from data.

Based on three different methodologies (DTs, neural networks and multinomial logit model), work travel mode choice was modelled by Xie et al. (2003). Data from San Francisco (California) were used and a C4.5 algorithm was applied for developing the tree. Five different alternative modes were proposed in the target variable (single-occupancy vehicle, carpool, transit, bicycle and walk) and two type of variables, individual and household sociodemographic attributes and trip level-of-service attributes, were used as explanatory variables. The results showed that the three models identified the same attributes as the most significant for the travel choice. These attributes were household vehicle number, license possession, travel time and out of-pocket cost. The results of the three models were compared and they indicated that the two data mining models offered better performances than the multinomial logit model. The DT obtained the highest estimation efficiency, while the neural networks reached a superior prediction performance in most cases. However, the interpretability in DTs is more explicit, for this reason, when the purpose of the investigation is policy making, DTs are preferred to neural networks, in order to know the reasons for a choice and to identify the most important variables.

One year after, Karlaftis (2004) used a multivariate DT for mode choice predictions. They used three different data sets for testing the methodology. The first data set concerning interurban mode choice (between Sydney,Melbourne and

New South Wales, Australia), and the second and third data set concern to commuters mode choice, one developed in Athens (Greece) and the other one in Las Condes (Chile). A cross validation test was used to evaluate the accuracy of the predictions, and the importance of the variables was also calculated, indicating effectively the most influential variables in choosing a mode. For the interurban case, four classes defined the target variable (air, train, bus, auto), and for the commuters cases, five classes (bus, bus/metro, metro, auto/metro, auto) and nine classes (auto driver, auto passengers, taxi, metro, bus, auto driver/metro, auto passenger/metro, taxi/metro, bus/metro) were defined respectively. The results indicated that the models performed successfully.

Differences in the predictive performance of DTs when it is considered a full model or a trimmed model were investigated by Moons et al. (2005). They generated various DTs in order to model different aspects of the activity-travel behavior (choice-facet level, activity-pattern level and trip-matrix level). The trees generated were grouped into two categories, first group for those models that predict the target variables using all the descriptive factors as independent variables, and second one, considering only the descriptive factors that were deduced as relevant in the previous tree for generate the tree. The results indicated that significantly smaller decision trees predicted well the different aspects of activity-travel behavior, and did not lose very much in the predictive power. In a previous paper (Moons et al, 2001), they also examined the influence of irrelevant attributes on the performance of the decision tree.

A new hybrid approach, combining rule-based of the conventional DTs and parametric discrete choice model, was introduced by Arentze and Timmermans (2007). The Parametric Action Decision Tree uses the tree to classify cases and the logit model to determine choice probabilities. It combines the specific strengths of rule-based models and parametric models of discrete choice, and overcomes the weakness of rule-based models of being less sensitive to continuous variation of level-of-service attributes of a transport system or land-use system. Two experiments were carried out with this new methodology: one for determine, for each individual and day, whether or not a work activity is included in the schedule; and the second model determines for each work activity the transport mode used for the trip. For the first model 21 independent attributes were used, and 39 attributes were used for the second model. The attributes describe characteristics of the household, person, study area (accessibility of locations), and facets of the activity pattern.

Recently, another study has been performed by Peng and Luan (2011) in order to deal with the traffic modal choice. The C4.5 algorithm was used for generating the tree and extract useful decision rules. Walking, bicycle, bus, taxi and private car were the different class values for the target variable (output of the model), and several factors, such as age, income, ownership of private car, trip time, trip purpose and trip distance, were used as input variables. 80% of the sample was used as the training data, and the rest as the testing data. The results showed that the model achieved a good accuracy rate and that the attribute used as the splitter of the root node was being ownership of a private car. Authors remarked that the rules obtained by the DT methodology were appropriate for forecasting the residents' traffic modal choice, and provided useful information for an accurate and reasonable traffic planning and design.

2.2.3.2. Others applications in transport engineering

Applications of DTs in transport engineering are not limited to the scope previously mentioned. Many other applications in traffic forecasting (Washington and Wolf, 1997), vehicle emissions (Washington et al., 1997; Wolf et al. 1998; Hallmark et al., 2002), road safety (Chang and Wang, 2006; Kuhnert et al. 2000; Pakgohar et al., 2011; Kashani and Mohaymany, 2011; de Oña et al., 2013) or even for classification of transit modes (Hodgson and Potter, 2010) have been identified. The following is a brief description of some of these applications.

One of the first contributions carried out in the field of traffic forecasting was developed by Washington and Wolf (1997). They used a hierarchical tree-based regression (HTBR) approach to forecast trip generation. The purpose of their study was to predict the number of daily automobile trips made by households for all trip purposes. Data from Michigan area were used and seven independent variables were considered (income, number of cars, number of adults, number of 16 to 18 years old, number of 5 to 15 years old, number of females and number of males). The results of this model were compared with the ones obtained using a classical ordinary least squares regression. Authors declared that ordinary least squares regression models were more intuitive and easy to interpreter than the HTBR, however HTBR could treat multicollinearity problems, does not need to define a functional form and could treat better non-additive and nonlinear behaviour.

The same year, Washington et al. (1997) applied the HTBR model in other field. The objective of this work was to determine modal correction factors for motor vehicle emissions. Based on 4,800 vehicle emissions tests, they obtained a satisfactory predictive accuracy in the model, and overcame the limitations that had the classical ordinary least squares regressions, as they have declared in their parallel work. The results showed that high- and normal-emitting vehicles were sensitive to different operational and vehicle specific factors, and the most influenced variables were identified. These factors were the changes in power requirements, the idle activity, the positive kinetic energy, the vehicle model year, the engine size, etc.

Likewise, the classification of vehicles into high and normal motor emitter based on hot-stabilized emissions was developed by Wolf et al. (1998). They used the

HTBR model as well Washington et al. (1997) and Washington and Wolf (1997), and a large number of vehicle and technology attributes on emitter status for separating the vehicles into homogeneous emitter categories. These authors pointed out the wide amount of advantages of this methodology, such as it was flexible with respect to the number of classes and types of variables used, it considered the influence of a large number of variables, and it ensured a good classification of the class variable which allowed developing separate emission rate models.

The problem of identifying geometric and operational roadway characteristics that influenced vehicle activity in order to use them as input in modal emission models was studied by Hallmark et al. (2002). They correlated the emission rates to specific ranges of activity using a HTBR methodology. A set of five models, every one using different dependent variable, were developed for various distance segment analyzed (segments created from the initial point on a queue to the traffic control in a intersection, because they indicated that vehicle activity was relatively homogeneous over certain length). The roadway and intersection geometric and operational factors and vehicle characteristics were used as independent variables of the trees; and the dependent variables were the percentage of vehicle activity spent in the specific operation modes (related with the acceleration, deceleration, speed and the inertial power surrogate). Regression tree analysis was used to identify the independent variables that had the most explanatory power in describing on-road vehicle activity. The authors indicated that this methodology was satisfactory in identifying the on-road geometric and operational variables influencing vehicle activity.

Kuhnert et al. (2000) developed a DT, a multivariate adaptive regression splines (MARS) and a logistic regression model for analyzing epidemiological case-control study of injuries resulting from motor vehicle accidents. The DT was built by a CART algorithm. The results indicated that the non-parametric techniques (CART and MARS) provided more informative and attractive models and the outcomes could be displayed graphically. However, the high complexity of MARS model converted it in a non choice as a modeling tool.

The relationships between crash severity and several characteristics related to drivers, vehicles, roads and the environment was studied by Chang and Wang (2006). Data from accidents in Tapei area were collected in 2001 and a CART model was developed. The target variable was categorized into three classes (fatality, injury and no-injury) according to the level of injury to the worst-injured occupant. The findings showed good overall predictions of the tree and this work demonstrated that CART analysis is an appropriate methodology for analyzing injury severity in traffic accidents.

Based on a DT (CART algorithm) and a Multinomial Logistic Regression, the specific influence of driver' characteristics into crash severity of an accident was investigated by Pakgohar et al. (2011). The results of the two methodologies were compared and revealed that CART model obtained more precise predictions and were also simpler and easier to interpret.

Kashani and Mohaymany (2011) used CART to identify the main factors that affect the injury severity of vehicle occupants involved in crashes on Iran roads. Due to the large amount of data they were going to treat, authors indicated that data mining techniques were highly suitable for this purpose, because they are able to discover meaningful models and patterns when great quantity of data are available. They discovered complex relationship between variables and found the most important variables between the vast amount of independent variables considered.

Also, the main factors affecting crash severity have been identified by de Oña et al. (2013). They used decision trees and compared the results obtained by different algorithms (ID3, C4.5 and CART). Their findings indicated that ID3 was the method that provided the worst results, while CART and C4.5 algorithms showed certain similarities in the structure of the tree and in the precision and parameters analyzed, being the difference in their improvement not significant.

The problem of classifying and defining with accuracy different transit modes, was carried out by Hodgson and Potter (2010). Based on various example systems (light rail of Manchester, trolley-bus of Seattle, guided-bus of Leeds, and so on), they applied a DT for explaining the most significant distinguishing characteristics between light rapid transit modes, such as guided-bus, trolley-bus, light rail and tram-train. Authors pointed out that current definitions used to describe transit systems are inconsistent and at times misleading. Characteristics related with vehicle capacity, on-street running and vehicle guidance were used as independent variables of the model. The findings provided a solid definition about the different transit modes that allows transport promoters and operators to have a consistent basis of reference when comparing and specifying rapid transit systems.

2.2.3.3. Applications of Decision Trees related to quality of service

In the field of service quality, DTs is a new technique that has not been used for many authors. There are some recent DTs applications that have used this technique for investigating some specific aspects related with customer satisfaction or service quality (Huang and Hsueh, 2010; Wong and Chung, 2007), however, there is no application to analyze quality of service for public transport operation, except the recently publications of the author of this Ph. D. thesis (de Oña et al. 2012a; de Oña et al. 2012b).

In 2007, Wong and Chung were the first (to our knowledge) to apply DTs for analyzing valuable and non-valuable passengers of an airline service, using service quality attributes. In their study they used a C5.0 DT and a cross-validation testing in order to classify the passengers of a Taiwanese domestic airline considering demographic profiles, travel behaviors and perceptions of service quality as independent variables.

Their results indicated that the model achieved reliable results and highlighted some of the key factors that are important to valuable customers. Some of these factors were related with passengers' satisfaction towards some aspects of the service, such as Satisfaction of need for fulfillment, Satisfaction of overall airline image, Satisfaction of accurate boarding announcements and Satisfaction of prompt reservation service.

Based on consumption characteristics, firm selection behavior and satisfaction degree, the consumer behavior in a refurbishment industry was analyzed by Huang and Hsueh (2010). The relationship between these concepts was studied using DTs. Four different models were generated and successfully results were achieved. The first DT was built for classifying and predicting the consumers' attitudes towards brands, to understand how are customers' perceptions about the brands. The second one was used to analyze how marketing policies affect customers' selection behaviour. The third one identified the key elements influencing the customers' price preference. And the last one was for evaluating the overall service quality satisfaction on the refurbishment industry. To our knowledge, this has been the first time that a tree model has been developed for predicting overall service quality in an industry. In this model four classes of the target variable were considered (very bad, bad, good and very good) and 22 items were used as independent variables (the 22 items described in the SERVPERF scale). Finally, in order to find out more applicable association rules, the four classes considered in the target variable were reduced into two (Bad and Good). The findings of this analysis showed that the equipment, real-time services and reputations of refurbishment firms were crucial key factors in SQ. The results of this study indicated that the reliability and applicability of these models was high, and that the decision rules extracted provided useful information to decision-makers in order to formulating business operations and marketing packages.

2.2.3.4. Summary

DTs methodology has been applied in different fields of transport engineering. Most of these studies defended the suitability of this approach over other classical regression models more popular in these research contexts. However, this methodology is still a novel technique for analyzing SQ, and particularly, SQ in public transportation. Nevertheless, DTs seem to be an adequate technique for this purpose due to the vast amount of advantages highlighted by different researchers.

Most part of the authors that have applied DTs in transportation choice behaviour point out that these models performs satisfactorily and achieve better and more interpretable results than conventional regression models widely employed for this purpose (Lee et al., 2010; Thill and Wheeler, 2000; Wets et al., 2000; Yamamoto et al., 2002; Xie et al., 2003). In fact, Lee et al. (2010) pointed out the important limitations of logistic regression models, as they present difficulties in interpret the coefficients, and also into assure that the developed model includes all relevant variables and excludes all irrelevant variables. Moreover, others DTs' advantages were highlighted among different authors, such as DTs facilitate the determination of the relationships between the explanatory variables and the choice (Yamamoto et al., 2002), the ability of DTs to uncover the structure of large and complex databases and to build this knowledge into operational transportation decision support systems (Thill and Wheeler, 2000), the advantage of DTs of being theory-free, not having predefined utility forms (as logit models) and establishing their structure derived from the analyzed data (Wets et al., 2000), the ability of extracting useful decision rules providing useful information for formulating adequate traffic planning and design strategies (Peng and Luan, 2011), and so on.

In the same way, the researchers working in other research fields of transport engineering (traffic forecasting, vehicle emissions, road safety, and so on), also stated that DTs overcame the limitations that had the conventional regressions models and achieved better performance (Kuhnert et al, 2000; Pakgohar et al., 2011; Washington et al. 1997; Washington and Wolf, 1997). Washington and Wolf (1997) remarked the suitability of this technique for analyzing traffic forecasting due to its ability to treat multicollinearity problems, does not need to define a functional form and could treat better non-additive and nonlinear behaviours. Kuhnert et al (2000) pointed out that non-parametric techniques provide more informative and attractive models, with outcomes displayed graphically. And Kashani and Mohaymany (2011) defended the high suitability of data mining techniques when large amount of data are going to be treat.

In the context of SQ, also the few researchers that applied DTs for analyzing any aspect of this concept affirmed that this methodology provided adequate and reliable results, and was a powerful tool for identifying the key factors affecting the dependent variable and extracting useful decision rules (Huang and Hsueh, 2010; Wong and Chung, 2007).

2.2.4. Advantages and disadvantages of using Decision Trees

After the brief literature review about DTs carried out in the previous section, it can be highlighted some of the advantages that this methodology presents. Some of which are the following:

- DTs have the ability to discover knowledge in large databases, identifying and explaining complex patterns in data, and complex interactions between the variables.

- DTs are non-parametric models with no assumptions about the relationship between the dependent and independent variables. In parametric models, when the model is misspecified, the estimated relationships between variables as well as the model predictions will be erroneous.

- The outcomes of the analysis are easy to understand due to the graphic representation afforded by the results, allowing non-statisticians or non-professional users to interpret the model. This is very important for public transport managers who would be able to interpret and manage this information directly.

- They permit to extract a set of "If-Then" Decision Rules, which provide useful and comprehensible information. Following a path from the root node to a terminal node, a decision rule is created and a prediction of the class variable is formulated.

- DTs can handle many explanatory variables and a large set of data. Both nominal and numeric variables can be used.

- Among the wide amount of variables considered, DTs can easily find the most important variables, and extract the weight of all the variables in the model.

- They can effectively handle multi-collinearity problems, which is one of the major drawbacks in regression models. In DTs the correlation between independent variables is not a concern.

- DTs are capable of handling outliers. Usually they are isolated into a node and do not create any effect on splitting. Eventually, they can even be pruned away. In parametric models, outliers can produce wrong estimations of the coefficients.

However, not all are advantages. This methodology also has some disadvantages, such as:

- The tree models are generally "unstable". Depending on the strategy followed for stratifying the sample in the training, validation and testing subsets, the structure and accuracy of the models generated could change.

- Moreover, unlike other parametric models, DTs do not provide a confidence interval or probability level to the splitters and predictions in the model.

2.3. Conclusions

As the preceding discussions indicate, SQ evaluation of public transportation poses formidable challenges: how to deal with such a complex, fuzzy and abstract concept as SQ; whether we should use performance perception only or also customers' expectations; which expectations should be considered (ideal, desired, adequate or tolerable quality); what is the relationship between SQ and satisfaction; how to identify the most relevant attributes that affect SQ; how to deal with subjective, qualitative and fuzzy data from surveys; possibility of using objective data (from transport companies) combined with subjective data for SQ analysis; customers satisfaction surveys limitations (maximum length of the survey, scale used, etc.); best ways to analyse heterogeneity; etc.

To deal with these challenges, innovative methodological approaches have been introduced in an attempt to improve the validity of the findings. From the beginning of 21st century in particular, this stream of methodological innovation has introduced some very exciting approaches that hold great promise in improving our understanding of the factors that affect the SQ in the public transport sector. One of these exciting and powerful techniques is the non-parametric data mining model denominated as Decision Trees.

DTs methodology has started to gain acceptance for applications in different fields of transport engineering (such as travel mode choice, spatial choice, vehicle emissions, road safety, etc) in the last decade. Moreover, DTs have been proven to be efficient for the analysis of different phenomenon in many different research fields (business administration, agriculture, industry, engineering and so on). Thus, given the characteristics of the data used to analyzed SQ, and the advantages of DTs that are beneficial for such characteristics, a utilization of DTs to the field of SQ in public transportation could be appropriate and could provide large benefits over other more conventional methodologies.

CHAPTER 3
OBJECTIVES

Chapter 3

OBJECTIVES

This Ph.D. thesis applies Decision Trees to the field of transit service quality modeling, given that Decision Trees have been proven to be able for dealing with complicated problems. Keeping in mind that many previous studies tried to employ different statistical techniques to analyze service quality in public transportation, the application of Decision Trees in the field of service quality in public transportation is new and generates a novel vision for solving some limitations of other popular statistical methods.

3.1. Principal objective

The main objective of this Ph. D. thesis is to validate that Decision Trees is an appropiate methodology for analyzing service quality in public transportation. The analysis of service quality brings practical value and useful information for transport planners and providers about the performance of the services from the point of view of passengers. This will permit them to design adequate marketing policies that promote increasing the use of public transport services and, therefore, a more sustainable mobility.

3.2. Specific objectives

For evaluating and defining the service quality in a public transport service, some specific objectives are followed in this doctoral thesis:

- Identify the most relevant variables influencing the overall service quality perceived by users and determine the weight of them in the model.

- Demostrate that passengers opinions and the variables most influencing their perceived service quality change when they are made to reflect on the attributes describing the service.

- Validate that the key factors influencing the overall SQ are different among market segments, and therefore, stratifying the sample of users in more homogeneous groups could help to reduce the heterogeneity present in their opinions, and therefore it would be helpful to planners of public transport services in order to draw up quality policies focus on groups of users with more uniform needs and perceptions about the public transport services (personalized marketing).

- Inquire in the problematic of stated importance rates about SQ attributes as a technique for determining the importance of the variables over the overall service quality.

- Research if the passengers' socioeconomic characteristics and travel habits variables have a high influence in the overall evaluation about the service.

CHAPTER 4

MATERIALS AND METHODS

Chapter 4

MATERIALS AND METHODS

In this chapter the different phases of the research work carried out in this Ph.D. thesis are exposed. Subsequently, a description of the methodology and the data used for achieving the proposed objectives are displayed.

4.1. Phases of the research work

The research work herein is divided in two differentiated experimental context. One for analyzing SQ of a bus public transport, and the other one for analyzing the SQ in a rail public transport. Then, the different analysis carried out are structured in the following steps:

4.1.1. Experimental context 1. Bus public transport

1. Firstly, the CART algorithm is used to build two different DTs for analyzing SQ in bus public transportation before and after passengers reflect on the characteristics that describe the service. CART methodology is described in section 4.2.1. and the data used correspond to 2007.

2. Using a specific stopping criterion the trees stop growing and subsequently the pruning process is performed. The stopping criterion is explained in section 4.2.2.

3. Using a 10-fold cross-validation technique and an Accuracy indicator, the performance of both DTs is compared. The 10-fold cross-validation and the evaluation indicator are detailed in sections 4.2.3 and 4.2.4.

4. The importance of the variables in both models is extracted by the Variable Importance Index (VIM), explained in section 4.2.5.

5. The variables importance derived from the models are compared among trees, and also with the importance rates stated by the users, in order to identify which are the key variables influencing passengers' opinions about the service before and after they reflect on the characteristics describing the service, and which are the main differences present in what they have declared.

6. Finally a set of decision rules are extracted from both DTs built. The construction and the quality evaluation of these decision rules is detailed in section 4.2.6.

7. Later, using another dataset of the bus public transport (data from 2008 to 2011), the overall market is split in 14 segments of passengers according to their socioeconomic characteristics and travel habits in order to diminish the heterogeneity present in the passengers' opinions

8. Using the CART algorithm, a global DT is built with the overall market and 14 different DTs are generated corresponding to the different market segments identified.

9. A comparison between the accuracy indicator evaluating the performance of the 15 DTs and the structure of the trees is carried out.

10. The most important variables for each segment of passengers are identified by the VIM and they are compared with the most important variables stated by users among segments.

4.1.2. Experimental context 2. Rail public transport

1. Firstly, the overall market of the Rail public transport is split in 13 segments of passengers according to their travel habits and according to the characteristics of the trip, in order to diminish the heterogeneity present in the passengers' opinions

2. CART algorithm is used to build 13 different DTs (one tree for each segment of the sample) and a global DT considering the whole dataset.

3. The accuracy indicator among the 14 DTs built is compared in order to evaluate the performance of the models.
4. Finally, using the VIM algorithm the importance of the variables is extracted and compare among market segments.

4.2. Methodology

The methodology used in this research work is described in this section. The CART algorithm used to build the DTs models (section 4.2.1), the stopping criterion used for stop the growing process (4.2.2), the evaluating technique applied for validating the model (section 4.2.3), the Evaluation Indicator used to evaluate the performance of the DTs (section 4.2.4), the algorithm used for discovering the importance of the variables (section 4.2.5) and the extraction and quality of the decision rules (section 4.2.6), all these concepts are explained in this section.

4.2.1. CART algorithm

The Classification and Regression Trees (CART) algorithm is a methodology for building Decision Trees developed by Breiman et al. in 1984. This algorithm has the ability of develop either types of tree, a Classification tree when the target variable is categorical, and a Regression tree when the target variable is continuous. Unlike other popular algorithms used also for building decision trees (ID3, C4.5, etc), CART methodology generates binary trees, splitting recursively the branches into two ways. The building-process in a regression tree is similar to the process in a classification tree, however, the splitting criteria used as well as the accuracy measure used are different among them.

The development of a CART model generally consists on three steps: The tree growing, the tree pruning and selecting the optimal tree. The first step is the tree growing. The principle behind the tree growing is to recursively partition the target variable to maximize "purity" in the two child nodes. Then, this process begins with all the data concentrated on the root node. On the basis of an independent variable (splitter), the root node is divided into two child nodes. The variable used as splitter is the one that creates the best homogeneity in the two child nodes. In fact, the data in each child node are more homogeneous than those in the upper parent node. The splitting process is applied recursively for each child node until all the data in the node are of the same class (the node is pure), their homogeneity cannot be improved, or an stopping criteria has been satisfied. In this case, terminal nodes are created.

The following example (Figure 4) shows a further explanation about the process of growing a CART tree. In this example the target variable is the Overall Service Quality perceived by the passengers of a public transport service. The target variable has three different classes: Poor, Fair and Good, and a Classification tree is built. The independent variables used as splitters were the Frequency of the service and the Information available about the service, both from the passengers' point of view. The CART algorithm splits the answer area into rectangular homogeneous areas, choosing the thresholds of the splitters (in this case denominated A and B) that higher homogeneity create in the child nodes. In this way, the tree model created (Figure 5) shows that if a passenger evaluates the Frequency of the service with a value minor than A, his overall evaluation about the service quality will be Poor (terminal node TN1). However, if his rate is higher than A, his overall evaluation about the service depends also on the variable Information. If Information is evaluated with a value minor than B, Then the target variable will be classified as Fair (terminal node TN2), but If the rate is higher than B, the target variable will be classified as Good (terminal node TN3).

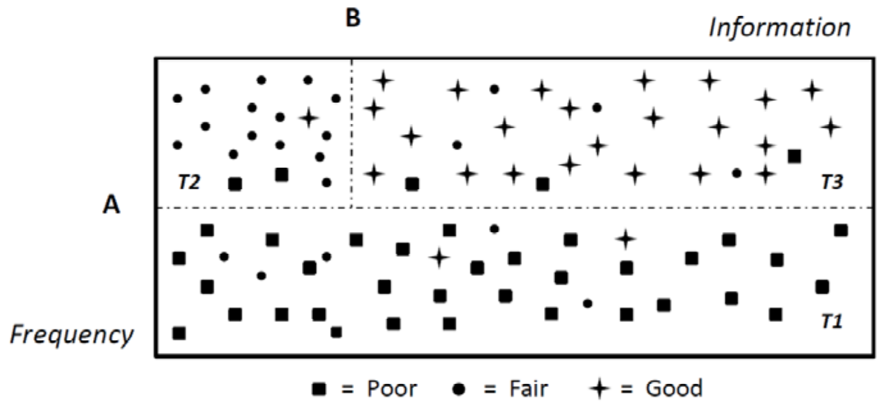


Figure 4. Example of a group of data in the answer area

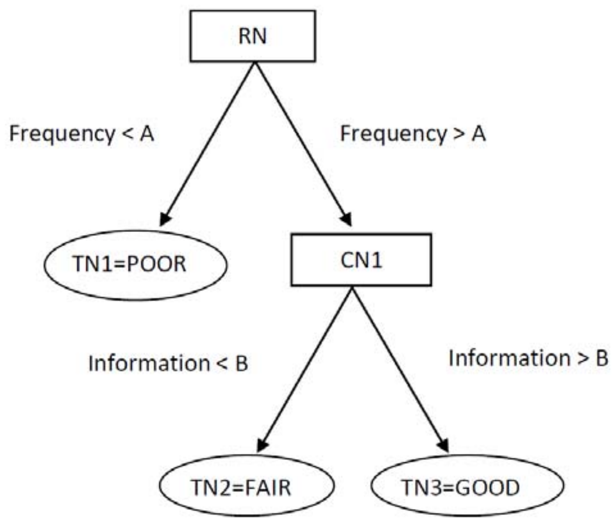


Figure 5. Example of CART model for the Overall Service Quality.

When the independent variables used as splitters are nominal and are composed of various categories, these categories are combined into the two associations that create the highest purity in the two child nodes. In numeric independent variables, the data is split according to a threshold of the variable used as splitter that also generates the best homogeneity in the child nodes. Normally, in this type of binary trees, the independent variables are retested across the path between the root node to a terminal node.

Then, during the tree growth, a set of candidate split rules is created, which consists of all possible splits for all variables included in the analysis. For nominal independet variables, supposing that the variable is defined by X different categories $C_1, C_2, \ldots, C_x$, the set of possible splits of this variable will be $2^{X-1}-1$. In numeric independent variables, the number of possible splits at a given node is one less than the number of its distincly observed values. These splits are then evaluated and ranked using a different splitting criteria for a Classification tree than for a Regression tree. The splitting criteria for the Classification tree is based on the Gini index, while the splitting criteria for the Regression tree is based on the Least Square (LS) error criterion.

Following this process and using different splitters a saturated tree is obtained. Normally, to develop a CART model, data are divided into two subsets, one for learning (or training) and the other for testing (or validation). The learning sample is used to split nodes, while the testing sample is used to compare the misclassification. The saturated tree is constructed from the learning data. Thus, the saturated tree provides the best fit for the data set which it is constructed from, but overfits the information contained within the data set. This overfitting does not help in accurately classifying another data set. This occurs because the tree fits idiosyncrasies and noise in the learning dataset, which are unlikely to occur with the same pattern in a different set of data. Now, to lessen the complexity of the saturated tree that overfits the learning data and to create simpler trees, the tree is "pruned" in the second step. This pruning is performed according to the Cost-Complexity algorithm for the Classificacion tree, or according to the Error-Complexity algorithm for the Regression tree. Based on the pruning algorithms, a set of pruned trees are created (T0, T1, ..., Tk). In the last step the optimal tree is selected, being the one with the smaller error rate in the testing subset. The error rate in classification trees is the missclassification cost, while for regression trees it is the mean square error.

Below, the different splitting criteria and algorithms used for pruning the Classification and Regression tree are explained.

4.2.1.1. Classification tree

A Classification tree is generated when the dependent variable or target variable is categorical. The development of the tree consist on the three steps explained previously: tree growing, tree pruning and optimal tree. In the tree growing step, the set of candidate split created for the recursively partition of the target variable, are evaluated and ranked using a splitting criteria based on the Gini index. The Gini index measures the impurity degree of a node in a tree. This impurity may be defined as follows (Kashani and Mohaymany, 2011):

$$\text{Gini (m)} = 1 - \sum_{j=1}^{J} p^2\,(j|m) \qquad (14)$$

Where *Gini (m)* is the impurity measure of a node *m*, *J* is the number of classes of the target variable, and $p(j|m)$ represents the conditional probability of an instance to belong to the class *j* when it is in the node *m*. This probability is defined as follows:

$$\mathbf{p(j|m)} = \frac{\mathbf{p(j,m)}}{\mathbf{p(m)}} \qquad \mathbf{p(j,m)} = \frac{\boldsymbol{\pi}\mathbf{(j)N_j(m)}}{\mathbf{N_j}} \qquad \mathbf{p(m)} = \sum_{\mathbf{j=1}}^{\mathbf{J}} \mathbf{p(j,m)} \qquad (15)$$

Where $\pi(j)$ is the prior probability of the class *j*, $N_j(m)$ is the number of instances of the class *j* in the node *m* and N_j is the number of instance of the class *j* in the root node. If a node is 'pure' (all the instances are of the same class), this measure (Eq. 14) will reach the minimum value equal to zero. On the other hand, the less homogeneous are the nodes, the value of the Gini index will be higher.

Then, the splitting criteria, denoted as the Gini Reduction criteria, measures the "worth" of each split in terms of its contribution toward maximizing the homogeneity of the child nodes through the resulting split. If a split results in splitting of one parent node into B branches, the "worth" of that split may be measured as follows:

$$\boldsymbol{\Delta}\mathbf{Gini(x_j, T)} = \mathbf{Gini(T)} - \sum_{\mathbf{b=1}}^{\mathbf{B}} \mathbf{P(b)\,Gini(b)} \qquad (16)$$

Where $\Delta Gini(x_j,T)$ represents the Gini Reduction measure at a parent node *T* which is split by a variable x_j. *Gini(T)* denotes the Gini index (impurity) of the parent node *T*, *P(b)* denotes the proportion of instances of the parent node assigned to the child node created with the branch b, and *Gini(b)* is the Gini index of the child node created with the branch b. So, considering the definition of the Gini Reduction criteria, a split resulting in more homogeneous branches will have a higher value of the "worth" or Gini Reduction.

Following the splitting criteria process until no more partitions can be created, the terminal nodes are created. At each of the terminal nodes it is predicted a class of the target variable, which is the one that have a higher representation of instances. The obtained tree will be a saturated tree that overfits the data. In order to create a tree that does not overfits the data the pruning process is performed in the second step. In the Classification tree this pruning is performed according to the Cost-Complexity algorithm, which is based on removing the branches that add little to the predictive value of the tree. This algorithm depends on a complexity parameter, denominated α. This parameter measures how much accuracy should be increased in the tree through an additional split to warrant the increased of its complexity. The Cost-Complexity algorithm is explained by Eq.17:

$$\boldsymbol{\alpha} = \frac{\boldsymbol{\varepsilon}\ (\mathbf{pruned}(\mathbf{T},\mathbf{t}),\mathbf{S}) - \boldsymbol{\varepsilon}(\mathbf{T},\mathbf{S})}{|\mathbf{leaves}(\mathbf{T})| - |\mathbf{leaves}(\mathbf{pruned}(\mathbf{T},\mathbf{t})|} \tag{17}$$

where $\varepsilon(\mathrm{T}, \mathrm{S})$ indicates the misclassification cost of the tree T over the sample S, $|leaves(T)|$ denotes the number of leaves in T and *pruned(T,t)* denotes the tree obtained by replacing the node t in T with a suitable leaf.

Thus, beginning from the last level (terminal nodes), the child nodes will be pruned away if the resulting change in the misclassification cost or classification error rate is less than α times the change in the tree complexity. The α parameter is gradually increased during the pruning process, subsequently more and more nodes are pruned away, and simpler and simpler trees are created. At the end of the pruning process, the result is a sequence of trees ({T0, T1,..., Tk},where T0 is the original tree before pruning and Tk is the root tree), and the relationship between the misclassification costs and tree complexity in terms of the number of terminal nodes (given in Figure 6).

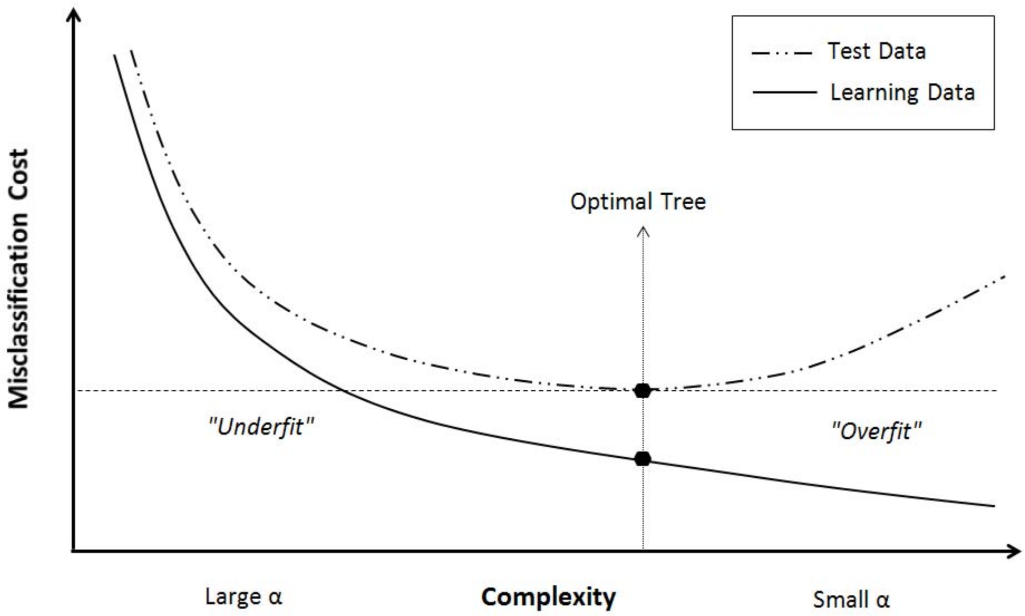


Figure 6. Relationship between tree complexity and misclassification costs

The last step is to select an optimal tree from the pruned trees (T0, T1, ..., Tk). The principle behind selecting the optimal tree is to find the correct complexity parameter α so that the information in the learning dataset is fit but not overfit. In general, finding this value for α requires look for a tree with respect to a measure of misclassification cost on the testing dataset (or an independent dataset). As shown in Figure 6, when the tree grows larger and larger, the misclassification cost for the learning data decreases monotonically, indicating that the saturated tree always gives the best fit to the learning data. On the other hand, in the misclassification cost for the testing data, first there is a decrease, and later an increased is observed, after reaching a minimum. Then, the optimal tree is the one that has the least misclassification cost for the test data. More detailed description of CART analysis and its applications can be found in Breiman et al. (1998).

4.2.1.2. Regression Tree

The Regression tree is produced when the target variable is continuous. The development is very similar to the classification tree, also consisting on the growing, pruning and selecting the optimal tree steps. In the first step the tree grows by recursive binary splits of the target variable into more homogeneous child nodes. The split is based on the heterogeneity of the data in terms of the variance of the target variable. The data in each child node are more homogeneous than those in the upper parent node. When terminal nodes are created, a value of the target variable is predicted at each of them. This value will be the average of the target variable of the data that reach this leaf. Moreover, at each terminal node, a measure of the standard deviation of the dependent variable is displayed.

In this case, the splitting criteria used for evaluate the set of candidate splitting rules is based on the Least Square (LS) error criterion. There is also an alternative method for using as splitting criteria for the regression trees in CART. This is the Least Absolute Deviation (LAD) function. However, the LS error criterion is the most common method. 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 reduce the impurity in the parent node is selected. This impurity can be measured as follows (Yohannes and Webb, 1999):

$$\mathrm{Err(t)} = \frac{1}{\mathrm{N_t}}\sum_{\mathrm{i=1}}^{\mathrm{N_t}}\left(\mathrm{y_{i(t)}} - \bar{\mathrm{y}}_{\mathrm{(t)}}\right)^2 \tag{18}$$

Where *Err(t)* is the impurity function in a node *t*, $y_{i(t)}$ are the individual values of the independent variable at the node *t*, $\bar{y}_{(t)}$ is the mean value of the target variable at the node *t* and N_t is the number of instances at the node *t*. In this way, the “worth” of a split may be measured as follows:

$$\Delta\mathrm{Err(s,t)} = \mathrm{Err(t)} - \frac{\mathrm{N_{tR}}}{\mathrm{N_t}} * \mathrm{Err(t_R)} - \frac{\mathrm{N_{tL}}}{\mathrm{N_t}}\mathrm{Err(t_L)} \tag{19}$$

Where, *ΔErr(s,t)* represents the Impurity Reduction measure at a parent node *t* with a *s* split, *Err(t)* is the impurity function in the parent node *t*, N_{tR} is the number of instances in the right child node, $Err(t_R)$ is the impurity function in the right child node, N_{tL} is the number of instances in the left child node and $Err(t_L)$ is the impurity function in the left child node. So the best split is the one that achieve the highest Impurity Reduction of the parent node.

Subsequently, after developing all the possible splits, the saturated tree that overfits the data is created. Then the tree is pruned in the second step. The pruning process uses the Error- Complexity algorithm, which works similar to the Cost-Complexity algorithm used in the Classification tree. In this case, the Error-Complexity algorithm will remove a branch when the resulting change in the Mean Square Error of the tree is less than α times the change in the tree complexity. In this case the Mean Square Error is used to measure the accuracy of the predictor. The optimal tree will be the one that achieve the lowest estimated error in the testing dataset. More detailed description of CART analysis and its applications can be found in Breiman et al. (1998).

In the research work carried out herein, the Classification Tree is the methodology used to build the DTs, due to the fact that the dependent variable analyzed is the SQ in different public transport services defined with three levels or categories of quality, denominated Poor, Fair and Good. So, next sections are focused on this type of DT.

4.2.2. Stopping criteria

In the tree growing process, the partition of the nodes is recursively developed until the node is pure, their homogeneity cannot be improved, or an stopping criteria has been satisfied. The most common stopping criteria used in decision trees are the following:

- A maximum tree depth has been reached
- The child nodes have less cases than the threshold established for parents nodes, then they cannot split and they become terminal nodes
- If the node were split, the number of cases in one or more child nodes would be less than the threshold established for child nodes, so it would not be split
- The best splitting criteria is not greater than a certain threshold established.

In this research work the unique stopping criterion used for building the DTs was a minimun number of instances for the child nodes. At each model built, at least a 1% of the sample should be represented in each child node.

4.2.3. Validation Technique: k-fold cross-validation

The best way for validate the predictive accuracy of a tree model relies on taking an independent dataset and run it down the tree to determine the proportion of instances misclassified. Normally, the complete dataset used for the analysis is divided into two subsets of data: the training data and the testing data. The training data is used for learning the sample, splitting the nodes and growing the tree, while the testing data is used to extract the accuracy rate of the model. It is common to hold out two-thirds of the data for training and use the remaining one-third for testing. The subsets of data (training and testing) should be chosen randomly and be representative, in the way that the proportion of data composing each class of the target variable in the full dataset should be represented in about the right proportion in the subsets created.

However, the k-fold cross-validation technique is an important statistical method used for validating the procedure for model building. This technique generates reliable results and ensures that the training and testing datasets are representative. Moreover, this is probably the validation method most practical in limited-data situations (Witten and Frank, 2005).

In general, in a k-fold cross-validation you decide on a fixed number of folds (k), or partitions of the data. It uses the whole dataset, and randomly divides the sample into k subsets in which the class is represented in approximately the same proportions as in the full dataset. Sequentially, each subset is used as a testing set, for the tree model generated by the remainder k-1 subsets that have been used for training. That is to use k-1 folds for training and one fold for testing, and repeat the procedure k times so that, at the end, every instance has been used exactly once for testing (see Figure 7). Thus, different k models are obtained, in which the accuracy of the classification or the error rate is calculated. Finally the k accuracy indicators are averaged to yield an overall accuracy value.

In CART methodology, when the pruning process is developed, a set of pruned trees is created for each k model obtained, varying the value of the parameter α. In order to select the optimal tree among the pruned trees, a finite set of candidate values for α is identified. The average error rate (ε_m) over the k folds when $\alpha = \alpha_m$ is calculated. Among the different candidate values of α established, the α_m choosed for pruning the original tree built with the entire dataset will be the one that minimizes ε_m. Then, the test trees built during the cross-validation process are used only to find the optimal tree.

The amazing fact on which cross validation is based is that the average accuracy from the k built models, is an excellent estimate of the performance of the original model produced using the entire dataset.

Extensive tests on numerous datasets, with different learning techniques, have shown that 10 is about the right number of folds to get the best estimate accuracy (Witten and Frank, 2005). For this reason in this research work we have used a 10-fold cross validation.

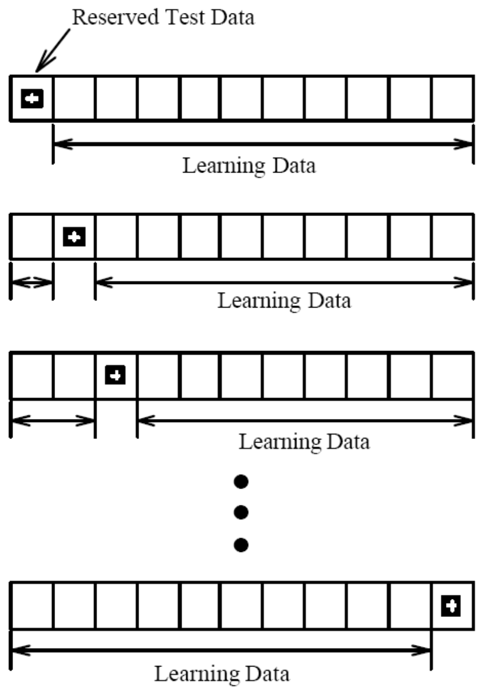


Figure 7. k-fold cross-validation procedure (Source: Lewis, 2000)

4.2.4. Evaluation indicator

The performance of the model is evaluated by the Accuracy rate. The Accuracy rate reflects the predictive power of the classification model, measuring the overall performance of the model. In order to define this index two concepts should be defined:

- *TP_i: is a True Positive.* Instances observed to be from the class *i* and are classified (predicted) correctly to belong to the class *i*.

- *FN_i: is a False Negative.* Instances observed to be from the class *i* but are classified (predicted) incorrectly to belong to other class ≠ *i*.

The Accuracy rate defines the proportion of instances that are correctly classified by the classifier of the method. To obtain this indicator an independent dataset that played no part in the formation of the classifier is needed. If a k-fold Cross-Validation technique is used, the predicted accuracy will be the average accuracy from the k built models.

Then, for a target variable that considers *I* different classes, which can be defined as C_i (i=1, 2,..., I), the Accuracy is defined by the following equation:

$$\text{Accuracy} = \frac{\sum_{i=1}^{I} TP_i}{\sum_{i=1}^{I}(TP_i + FN_i)} * 100\% \tag{20}$$

The more predictive power achieves the classifier, the more high accuracy indicator will have.

4.2.5. The importance of the variables

One of the most valuable outcomes provided by CART analysis is the value of the standardized importance of independent variables. It can be obtained by the Variable Importance Index (VIM), which reflects the impact of the predictor variables on the model. The information is obtained for all the independent variables, making it easy to find which ones are the most important. Therefore, the relative importance of a variable x_j is defined in the following equation (Kashani and Mohaymany, 2011):

$$\boldsymbol{VIM\ (x_j) = \sum_{t=1}^{T} \frac{n_t}{N} \Delta Gini(x_j, t)} \tag{21}$$

Where,

$\Delta \text{Gini}(x_j, t)$: is the Gini reduction at a node *t* that is achieved by splitting by the variable *xj*,

$\boldsymbol{\frac{n_t}{N}}$: is the proportion of the observations in the dataset that belong to node *t*,

T: is the total number of nodes and *N* is the total number of observations.

4.2.6. Extracting Decision Rules

Decision Rules is a very useful information that is directly extracted from the decision trees. Every terminal node created in a tree generates a rule. These rules comprise a set of antecedent conditions created on the path from the root node to the leaf, and the consequent of the rule is the predicted class of the target variable at this leaf (Witten and Frank, 2005). Therefore, each one of the paths through the tree produces a different rule.

Decision rules generally take the form of “If-Then” statements, where *If* contains a set of conditions that are a conjunction of attributes tests since the root node to a leaf following the structure of the tree, and *Then* contains the consequent of the compliance of these conditions as one state of the class variable. The set of conditions found in the path from the root node to a leaf are determined from the different independent variables used as splitter which, when they are nominal, the associations of the categories establish the condition, while, when they are numerical, the threshold established for the partition determines the condition.

Thus, the part IF is called the rule antecedent and the part THEN is called the rule consequent. Its logical conditional structure permits extract potencial useful and ready to use information from the decision trees.

For example, the decision tree built in the previous example (Figure 5) generates the following decision rules:

- IF (Frequency < A) THEN (Overall SQ = Poor)
- IF (Frequency > A AND Information < B) THEN (Overall SQ = Fair)
- IF (Frequency > A AND Information > B) THEN (Overall SQ = Good)

The quality of the extracted decision rules is evaluated by two indices: the *Support index* and the *Confidence index*. The Support index of a decision rule is the number of instances that are correctly predicted in a terminal node from the entire sample:

$$Support = \frac{\mathrm{n_{Ti}}}{\mathrm{N}} * 100\% \tag{22}$$

Where,

n_{Ti}: represents the number of cases that are correctly predicted in a terminal node T, and

N: is the total number of cases that compose the whole sample.

The Confidence index is the number of instances correctly predicted among the proportion of instance reaching the terminal node.

$$Confidence = \frac{\mathrm{n_{Ti}}}{\mathrm{n_T}} * 100\% \tag{23}$$

Where,

n_T: is the number of intances that reach the terminal node T

High levels of Support and Confidence indeces are desirable for obtaining high quality rules. Minimum indices are sometimes established in some research works. However, these values depend on the nature of the data (if they are balanced or imbalanced) and also on the objective of the research, because sometimes decision rules created for rare events or under-representated classes of the target variable are the most useful and interesting rules, and in general, they obtaine low values for both indices (support and confidence). For example, transport planners will have high interest in decision rules created for poor perceptions of passengers about the service, which usually are defined by a little subset of passengers. For this reason, in this research work, no threshold are established, and only the quality of the decision rule is analyzed.

4.3. Data

In this research work two different experimental contexts in two different countries have been analyzed. Data from two public transport services have been used for validating that DTs, and in particular the CART methodology, is a suitable technique for evaluating SQ in public transportation. A metropolitan bus transit service and a suburban rail transit service have been studied. The data used in this Ph.D. thesis were gathered in various CSSs carried out in both services in order to know the passengers' opinions about the service.

In the following sections (4.3.1. and 4.3.2), the developed surveys and the collected data are explained.

4.3.1. Bus Public Transport

The bus public transport analyzed in this research work corresponds to the metropolitan PT service of the city of Granada (Spain). Granada is a medium-sized city in southern Spain with a population of 523,845 in the metropolitan area. A Granada Area Transport Consortium was created in 2003 to coordinate transit bus service management in the Metropolitan Area. The PT service in the metropolitan area carries more than 10 million passengers every year. It is provided by a bus system in which 15 bus companies operate in 18 independent transport corridors linking the metropolitan municipalities with the centre of the city of Granada.

The lines network is established by a radial structure focused on two central areas of the city of Granada, one in the north and the other one in the south of the city, and extending in all directions (corridors) to the rest of the urban agglomeration. Owing to the fact that Granada municipality population represents almost half of the total population in the metropolitan area, and also the main trip generators centers are located there (such as administrative centers, health centers, educational and commercial centers), it has produced that the structure of the transport system has been generated with this shape.

Since 2003, various improvements has been implemented by the Transport Consortium in the metropolitan transport system. These improvements involve establishing an Integrated Fare System, increasing the number of service a day, creating new services in areas of urban growth, etc. The main interventions were made in the first years in which the Consortium was established, between 2003 and 2007. In the last years the changes have not been very significant.

Moreover, in 2006, the Transport Consortium conducted the first CSS to evaluate SQ in the Granada Metropolitan PT system. Since this year, it has developed an annual CSS to analyze changes in the perceived SQ of the passengers. Each year more than a thousand users are interviewed in the months of March or April (see

Table 5). The surveys are collected through a face-to-face questionnaire proposed to the users at the main bus stops of the lines.

Table 5. CSSs conducted by the Transport Consortium of Granada

YEAR	DATE	SURVEYS
2006	From 21/3 to 24/3	1071
2007	From 13/3 to 17/3	1200
2008	From 10/3 to 15/3	1278
2009	From 30/3 to 2/4	1297
2010	From 23/3 to 25/3	1292
2011	From 5/4 to 25/4	1625
2012	From 5/4 to 25/4	1729

The first survey developed in 2006 was very different from the subsequent ones, regarding the scale used (4-point likert scale) and the items asked. In 2007 the overall evaluation about the service quality was asked twice during the survey, once before users have reflected about the attributes describing the service, and the other one later. In addition, passengers were asked to state the importance and perceptions of the service quality attributes in a 11-point likert scale. Since 2008, the survey also changed from the previous one, but it remained equal in the later years (from 2008 to 2012). Since this year, the importance of the attributes was asked by ranking the three most important attributes.

The data used for this research work are those collected in the CSSs carried out between 2007 and 2011, because the survey used in 2006 was subtantially different from the survey collected in others years, and because the data from 2012 were not available until the end of 2012. The analysis developed in the bus metropolitan service of Granada is separated into two study cases. In order to analyzed segments of users in the second study case, it is neccesary to join data from different years with the purpose of obtaine enough data at each subsample created. So, the data collected in the period of time in which no substantial improvements of the service have been made, and also the questionnaire has not change were joined. Therefore, the data were split into two different periods of time. On the one hand the data from 2007, and on the other hand the data collected from 2008 to 2011 in which no modifications of the questionnaire were carried out.

4.3.1.1. CSS conducted in 2007

The CSS conducted in 2007 was structured into two main sections. The first section gathered general information (e.g. operator, line, time of the interview, origin/destination), demographic characteristics (e.g. sex, age, occupation) and travel habits (e.g. reason for travelling, frequency of use, type of ticket, availability of a private vehicle, complementary modes used for access to/ moves from the bus stop).

1,200 interviews were collected through a face-to-face questionnaire proposed to the users at the bus stops The sample is characterized by (see Table 6) a higher number of females than males (66.3% vs. 33.7%). More than a half of the users are aged between 18 and 30 years old (56.1%), 34.4% between 31 and 60, and only the remaining 9.5% are older than 60 years old. Employees (37.8%) and students (31.1%) constitute more than two thirds of the sample, while pensioners, unemployed, housewives and others represent the other third (31.1%). Almost half of the passengers use the service daily (46.9%), and 38.1% take the bus with a weekly frequency. Only 15% of the sample travel occasionally. The type of ticket used by the passengers is almost equally spread between the consortium pass (48%) and the standard ticket (41.3%). Only a little part of the sample uses the Senior citizen pass (6.5%) or another type of ticket (4.2%).

Table 6. Sample characteristics (year 2007 in the metropolitan bus transit service)

Characteristics	*Statistics*
1.Gender	Male (33.7%), female (66.3%)
2.Age	18-30 (56.1%), 31-60 (34.4%), > 60 year-olds (9.5%)
3.Occupation	Employees (37.8%), students (31.1%), others (31.1%)
4. Frequency of journey	Daily (46.9%), weekly (38.1%), occasionally (15%)
5. Type of ticket	Consortium pass (48%), Standard ticket (41.3%), Senior citizen pass (6.5%), other (4.2%)
6. Travel reason	Work (26.1%), studies (19.5%), doctor (13.4%), others (41%)
7. Private vehicle available	Yes (38.2%), No (61.8%)
8. Complementary modes from origin to bus stop	On foot (78.3%), urban bus (16%), other modes (5.8%)
9. Complementary modes from bus stop to destination	On foot (94.1%), urban bus (1.9%), other modes (4.0%)

Concerning the purpose of the trip, passengers have different reasons for travelling. For 26.1% the main reason is reaching the work place. Another important group (19.5%) travels for studying, and 13.4% of the passengers for going to the doctor. The rest of the sample (41%) stated that they travel for holidays, shopping or others personal activities. Out of all the surveyed passengers, only 38.2% could use a private vehicle for that trip. Also information about the complementary transport modes used by passengers for accessing to/moving from the bus stop was collected. Particularly, most of the sample accesses to the bus stops on foot (78.3%), 16% takes the urban bus, and the rest of the sample (only 5.8%) uses other modes (e.g their own car, motorbike, bicycle, etc). Likewise, travelling on foot is also the complementary mode mostly used for moving from the stops to the destination (94.1%), having the other transport modes a very low percentage.

The second section of the questionnaire focuses on the users' opinions about the service. This part is also divided in 3 main sub-parts: Part A, according to which passengers were asked to state the importance of each of the attributes describing the service, Part B, referred to the perceptions about the quality of each of these attributes, and Part C, collecting a global evaluation of the service quality. This last question was asked twice during the survey: once at the beginning of the second section (Previous Evaluation) and again at the end of the questionnaire (Later Evaluation), when the passengers were made to reflect on the attributes describing the service.

The service attributes considered in the survey are the following: frequency of the runs (Frequency), punctuality of the runs (Punctuality), speed of the trip (Speed), proximity of the stops to/from the origin/destination (Proximity), fare of the ticket (Fare), cleanliness of the vehicle (Cleanliness), space in the vehicle (Space), temperature in the vehicle (Temperature), available information (Information), safety on board (Safety), courtesy or kindness of the personnel (Courtesy), and easiness to get on/off the bus (Accessibility).

An 11-point Likert scale, from 0 to 10, was used for measuring importance and perceptions with the attributes and for the Later Evaluation of the overall service quality, while a 5-point semantic scale (Very poor, Poor, Fair, Good and Very good) was used for the Previous Evaluation. Table 7 shows the structure of the second section of the survey, the attributes assessed, the scale used to measure the attributes, and the average rate and standard deviation for the importance and satisfaction rates stated by the users.

According to the importance of the attributes, the judgments of the passengers show similar and low values of the standard deviation among the attributes (<1.8), therefore, their opinions are quite homogeneous. Punctuality, Frequency and Safety obtained the highest average rate, while Information, Space and Proximity the lowest ones. However, all the attributes are considered highly important, with average values comprised on the top of the scale (between 8.60 and 9.14). Furthermore, little variation exists among these mean values (only 0.5 points of variation among all the attributes). This insufficient differentiation among the evaluations makes difficult to identify which are the key factors really affecting the OSQ.

On the contrary, the judgments of the perceptions are more heterogeneous among the users, with values of the standard deviation higher than the values obtained in the Importance rates (higher than 1.8 and lower than 2.56). The attribute judged as the most heterogeneous is Fare, which is also the attribute with the lowest average rate (6.06). The average rates of the perceptions are lower than the mean values of the importance rates. They are concentrated in a range from 6 to 8. Nonetheless, these values are quite good, because all the attributes are perceived at least with an adequate quality (>6), and some of them with a quite good quality (>7). The attributes characterized by the highest levels of quality were Courtesy, Safety and Temperature.

Table 7. Section 2 of the questionnaire (year 2007)

Parts	Variables		Average Rate	Standard Deviation
A. Importance of the attributes	Item1	Frequency	9.03	1.54
	Item2	Punctuality	9.14	1.44
	Item3	Speed	8.72	1.70
	Item4	Proximity	8.68	1.77
	Item5	Fare	8.72	1.80
	Item6	Cleanliness	8.85	1.47
	Item7	Space	8.66	1.71
	Item8	Temperature	8.71	1.62
	Item9	Information	8.60	1.72
	Item10	Safety	8.98	1.52
	Item11	Courtesy	8.74	1.75
	Item12	Accessibility	8.85	1.78
B. Perceptions of the attributes	Item1	Frequency	6.80	2.53
	Item2	Punctuality	7.28	2.30
	Item3	Speed	7.23	1.95
	Item4	Proximity	7.34	2.17
	Item5	Fare	6.06	2.56
	Item6	Cleanliness	7.43	1.81
	Item7	Space	7.14	2.01
	Item8	Temperature	7.37	1.95
	Item9	Information	6.62	2.42
	Item10	Safety	7.65	1.96
	Item11	Courtesy	7.94	1.80
	Item12	Accesibility	6.75	2.44
C. Overall SQ	Item13	Previous Evaluation*	3.52	0.83
	Item14	Later Evaluation	7.07	1.58

*** The Importance and Perceptions of the attributes, as well as the Later Evaluation about the Overall SQ were measured in a 11-point likert scale (from 0 to 10), while the Previous Evaluation was measured in a 5-point semantic scale (from Very Poor to Very Good)**

By observing the average rates of the Previous and Later Evaluations, they show similar average values, with a value of 3.52 in the Previous Evaluation according to a 5-point scale (which is equivalent to a 6.30 in an 11-point scale) and a value of 7.07 in the Later Evaluation according to the 11-point scale. Then, users evaluate better the overall service quality when they have reflected on the different attributes characterizing the service.

4.3.1.2. CSSs conducted in the period 2008-2011

For the period comprised among 2008-2011, the CSSs were performed with the same questionnaire. It was also divided into two main sections. The first section was equal to the previous survey, gathering information related with general information aspects, socioeconomic characteristics, and travel habits (see Table 8).

Table 8. Sample characteristics (CSSs for the period 2008-2011 in the metropolitan bus transit service)

Characteristics	*Statistics*
1.Gender	Male (32%), female (68%)
2.Age	18-30 (49.4%), 31-60 (40.4%), > 60 year-olds (10.3%)
4. Frequency of journey	Frequent (78.3%), Sporadic(21.7%)
5. Type of ticket	Consortium pass (66.7%), Standard ticket (23.2%), Senior citizen pass (6.8%), other (3.3%)
6. Travel reason	Work (28%), studies (24.9%), doctor (11.5%), others (35.6%)
7. Private vehicle available	Yes (46.4%), No (53.6%)
8. Complementary modes from origin to bus stop	On foot (77.2%), vehicle (22.8%)
9. Complementary modes from bus stop to destination	On foot (95%), vehicle (5%)

The respondents of this period of time were characterized by being the majority of them female, with 2,493 (68%) female and 1,171 male (32%). Half of the respondents were age 18 to 30 (49.4%); 1,479 (40.4%) were age 31 to 60 and only 376 (10.3%) were older than 60. For 1,027 (28%) of the respondents the reason for travelling was occupation and for 911 (24.9%) the reason was studies. The rest of the respondents (47.1%) travelled for other reasons, such as doctor, shopping, holidays and so on. Most of the respondents used the service frequently (more than once a week), with 2,870 (78.3%) frequent passengers and with 794 sporadic passengers (21.7%). 2,445 respondents used the consortium pass (66.7%), as opposed to 850 (23.2%) who used the standard ticket, 249 (6.8%) or the senior citizen pass; 120 (3.3%) used some other type of ticket.

The second section of the survey is specifically about passengers' perception of several service characteristics. This part was also divided into 3 main sub-parts. In Part A, interviewers asked the passengers about their perception of performance with regards to 12 SQ factors, on a 11-point likert scale from 0 to 10. In Part B, they asked passengers to identify and rank the three most important SQ factors among the 12 factors describing the service. And finally, Part C was for asking passengers about the overall SQ perception based on a 5-point scale from 1 to 5.

The variables used to measure the perception of the SQ attributes were the same than in the 2007 CSS. They included frequency of the runs (Frequency), punctuality of the runs (Punctuality), speed of the trip (Speed), proximity of the stops to/from the origin/destination (Proximity), fare of the ticket (Fare), cleanliness of the vehicle (Cleanliness), space in the vehicle (Space), temperature in the vehicle (Temperature), available information (Information), safety on board (Safety), courtesy or kindness of the personnel (Courtesy), and easiness to get on/off the bus (Accessibility). Table 9 shows section 2 of the survey, displaying the average rates and standard deviation calculated from the performance perception rates expressed by the users with regards to the 12 SQ attributes, and also with regards to the Overall Evaluation.

Table 9. Section 2 of the questionnaire for the 2008 to 2011 period

Parts	Variables		Average Rate	Standard Deviation
A. Perceptions of the attributes[1)]	Item1	Frequency	6.08	2.51
	Item2	Punctuality	7.27	2.10
	Item3	Speed	6.97	2.09
	Item4	Proximity	7.10	2.21
	Item5	Fare	6.14	2.40
	Item6	Cleanliness	7.47	1.76
	Item7	Space	7.09	2.01
	Item8	Temperature	7.37	1.81
	Item9	Information	6.43	2.31
	Item1	Safety	7.59	1.85
	Item1	Courtesy	7.91	1.84
	Item1	Accessibility	7.28	2.03
B. Importance of the attributes[2)]	Item1-Item12			
C. Overall SQ[3)]	Item1	Overall	3.59	0.77

[1)] Using a 11-point Likert scale (from 0 to 10), [2)] using a Ranking scale (from 1 to 3) and [3)] using a 5-point Likert scale (from 1 to 5)

In general, the average SQ rates suggest that people are fairly satisfied with the service. All the attributes have an average rate higher than sufficient (>6). The service characteristics considered to have the highest SQ are Courtesy (7.91), Safety (7.59), Cleanliness (7.47) and Temperature (7.37). These four characteristics also have the lowest dispersions in their evaluation (all of them present a standard deviation lower than 1.85). On the other hand, the service characteristics with the lowest SQ but with the highest dispersion are Frequency (6.08), Fare (6.14) and Information (6.43).

By comparing the passengers' perceptions of the attributes in the data collected in 2007 and those collected among 2008 to 2011, it can be observed that, in general, they are very similar, although the tendency has been to have a poorer perception of the attributes describing the service. It might be explained by three different reasons: 1) the passengers have become more critics with the service, 2) the main interventions make by the transport consortium of Granada for improving the quality of the service were developed in the period of time between 2003 to 2007, and little interventions were made later and 3) the constructions of the metro started in April of 2007 (and they have still not finished) and have disturbed the ordinary performance of the service.

Nevertheless, there are some attributes that change the general tendency of the passengers' perceptions, such as the Fare and Cleanliness which suffered a slight improvement, or the Accesibility, which improvement was more significant (changing from an average rate of 6.75 in 2007 to an average rate of 7.28 for the period of 2008 to 2011). The reason for this could be due to the fact that the number of passengers that use the consortium card is higher every year, paying lower prices than those that use a standard ticket (from a 48% of the sample in 2007 to a 66.7% in 2008-2011). Moreover, the number of vehicles that are adapted to low mobility people, is growing every year.

The Overall Evaluation shows an average rate of 3.59 according to a 5-point Likert scale. So, the passengers' perceptions about the global SQ of the service provided is quite good. Moreover this value shows low dispersion among users (standard deviation equal to 0.77). By comparing the Overall Evaluation between 2007 and 2008-2011, the Later Evaluation in 2007 achieved higher average rates (7.07 in a 11-point scale) than the Overall Evaluation in the period 2008-2011, in which the obtained average value was of 3.59 in a 5-point scale (which is equivalent to 6.47 in a 11-point scale).

Table 10. Importance frequencies for the overall market

	Option 1*	**Option 2***	**Option 3***	**Sum**	**Overall (%)**
ACCESIBILITY	85	125	90	300	9.2%
CLEANLINESS	88	184	122	394	12.0%
COURTESY	115	149	123	387	11.8%
FARE	***603***	***534***	***468***	***1,605***	***49.1%***
FREQUENCY	***618***	***617***	***465***	***1,700***	***52.0%***
INFORMATION	159	80	120	359	11.0%
PROXIMITY	125	135	159	419	12.8%
PUNCTUALITY	***804***	***451***	***322***	***1,577***	***48.2%***
SAFETY	***369***	***329***	***288***	***986***	***30.1%***
SPACE	93	123	144	360	11.0%
SPEED	***173***	***306***	***314***	***793***	***24.2%***
TEMPERATURE	39	69	57	165	5.0%
n.a	393	562	992		
Num. observations	**3,271**	**3,102**	**2,672**		

*: number of times chosen as the attribute most important (Option 1), the second most important (Option 2) and the third most important (Option 3)

n.a.: not available

Table 10 shows the number of times that the passengers identified each attribute as being the most important (Option 1), the second most important (Option 2) or the third most important (Option 3). Users judged three attributes as very important (with a frequency of around 50%): Frequency, Fare and Punctuality. Two of them (frequency and fare) were also identified as attributes with the lowest SQ. Safety and Speed were also among the five most important attributes.

4.3.2. Rail Public Transport

The other public transport service analyzed in this Ph.D. thesis is a railway service operating in the North of Italy, and specifically in the city of Milan. This service is offered by 9 suburban lines connecting towns of the hinterland of Milan. The lines are used by about 200,000 passengers per day. The data were collected in a CSS conducted in the month of May 2012, when 7.333 users were interviewed. Face-to-face interviews were carried out on board during a whole week in a time slot between 6.00 a.m. and 10.00 p.m.

The questionnaire was structured into two main sections. Through the first section, data concerning general information (e.g. time period of the interview, train, line, station, and operator), socio-economic characteristics (e.g. gender, age, qualification, professional condition, and income), and travel habits (e.g. trip scope and frequency, and ticket) were collected.

The sample is made up more of females (see Table 11). Most of the passengers are aged between 16 and 25, and another fair chunk is represented by people aged between 26 and 40. The major part of the sampled people are students and employees. More than half of the sample obtained a diploma of a secondary school of second level, and almost 30% has a degree. More than one fifth of the sample doesn't give any kind of information about the income, while about 40.0% has not a fixed income; people stating their income mainly belong to a class between 1,001 and 1,500 Euros. Passengers travel by train mainly for reaching the place of work or study (73.2%). Most of the sample travels by train every day (61.9%), but about 24% of passengers occasionally travel. People mainly purchase a travel card (74.7%), and about 25% travel using a one-way ticket.

The second section is specific about passengers' perceptions of the used services; users expressed satisfaction and importance rates, on a 10-point likert scale from 1 to 10, about 27 service quality factors concerning safety, cleanliness, comfort, service, information, personnel and other, and also an overall satisfaction rate about the service using the same scale.

Table 11. Sample characteristics in the Rail public service

Characteristics	*Statistics*
1.Gender	Male (45.8%), female (54.2%)
2.Age	16-25 (40.5%), 26-40 (33.1%), 41-65 (22.4%), > 65 year-olds (4.0%)
3.Professional condition	employee (37.7%), manager (1.6%), entrepreneur (1.3%), freelancer (5.2%), self-employed worker (4.9%), unemployed (4.1%), student (38.1%), housewife (2.1%), pensioner (4.3%), other (0.7%)
4.Income level	<= 1,000 (9.6%), 1,001-1,500 (14.5%), 1,501-2,000 (7.4%), 2,001-3,000 (3.8%), 3,001-4,000 (1.3%), > 4,000 Euros (1.5%), no fixed income (39.5%), no answer (22.5%)
5.Qualification	Degree (28.9%), diploma of secondary school of second level (55.6%), diploma of secondary school of second level (14.0%), diploma of primary school (1.6%)
6.Scope of journey	Work (40.4%), studying (32.8%), bureaucratic activities (3.6%), personal activities (20.6%), tourism (2.6%)
7.Frequency of journey	Daily (61.9%), weekly (14.5%), occasionally (23.6%)
8.Ticket kind	One-way ticket (25.3%), travel card (74.7%)

The users judged most of the attributes as very important (showing an average rate of importance around 8 and 9); the attributes considered as the most important are the three attributes concerning travel safety, which also present little dispersion among users, showing homogeneous opinions (standard deviation <1.76). The attributes considered relatively less important are: bicycle transport on board (6.00), parking (7.04), info connection with public transport (7.33), and complaints (7.48). According to these attributes the judgments of the passengers are more heterogeneous with values of the standard deviation higher than 2.7 and lower than 3.37. The judgments of importance among passengers change from quite homogeneity to low homogeneity depending the attribute considered.

On the other hand, the satisfaction rates among users are more homogeneous than the importance rates, with standard deviation lower than 2.21. The average satisfaction rates suggest that people are not very satisfied with the service, in fact only nine attributes out of 27 have an average rate higher than the sufficiency (>6). The service characteristics considered as the most satisfying regard safety and personnel; all the other characteristics are judged as not satisfying. In addition, users consider the quality of the service on the whole as almost sufficient (average rate of satisfaction on the overall service of 5.81) (Table 12).

Table 12. Average and Standard Deviation of the Importance and Satisfaction rates

Service aspect	*Service quality attribute*	Import. Average	Satisf. Average	Import. St.Dev.	Satisf. St.Dev.
Safety	1. Travel Safety	8.98	7.43	1,76	1,91
	2. Personal Security on Board	9.01	6.76	1,70	2,06
	3. Personal Security at Station	9.00	6.48	1,76	2,14
Cleanliness	4. Cleanliness of Vehicles	8.43	5.32	1,85	2,11
	5. Cleanliness of Seats	8.48	5.22	1,82	2,12
	6. Cleanliness of Toilet Facilities	8.26	4.44	2,18	2,06
	7. Cleanliness of Stations	7.96	5.53	1,96	1,95
	8. Maintenance of Stations	7.82	5.49	2,04	1,94
Comfort	9. Crowding on Board	8.01	5.33	2,01	2,21
	10. Air-conditioning on Board	8.09	5.41	1,84	2,16
	11. Windows and Doors Working	7.95	5.74	2,04	2,06
Service	12. Fare/Service Ratio	8.39	5.17	1,97	2,10
	13. Frequency of Runs	8.41	6.12	1,89	1,97
	14. Punctuality of Runs	8.69	5.52	1,82	2,11
	15. Regularity of Runs	8.53	5.80	1,93	1,95
	16. Price Integration with PT	7.54	5.95	2,76	1,79
	17. Localization of Stations	7.84	6.65	2,30	1,69
Other	18. Parking	7.04	5.49	2,96	1,94
	19. Bicycle Transport on Board	6.00	6.03	3,37	1,50
	20. Facilities for Disabled	7.91	5.25	3,14	1,86
Information	21. Information at Stations	8.04	5.72	2,15	1,97
	22. Information on Board	8.00	5.45	2,15	1,98
	23. Complaints	7.48	5.17	2,70	1,75
	24. Info Connections with PT	7.33	5.26	2,78	1,75
Personnel	25. Courtesy and Competence on Board	7.92	6.67	2,08	1,77
	26. Ticket Inspection	7.68	6.20	2,26	2,03
	27. Courtesy and Competence in Station	7.91	6.38	2,14	1,88
	Overall service		5.81		1,62

CHAPTER 5

RESULTS AND DISCUSSION

Chapter 5

RESULTS AND DISCUSSION

In this chapter, the results obtained by constructing Decision Trees in different experimental contexts are presented. The software used to build the DTs was Weka (Witten and Frank, 2005), which is an open source freeware, available at: http://www.cs.waikato.ac.nz/ml/weka/.

Three different study cases are presented in this Chapter, and at each of them different objectives were persued.

Study Case 1: Decision Trees for the Pre-Evaluation and Post-Evaluation of the bus public transport. The main objectives followed were:

- To validate that Decision Trees is an appropiate methodology for analyzing service quality in public transportation.
- To demostrate that passengers overall evaluation about the quality of the service change before and after they are made to reflect on the attibutes defining the service. Moreover, to prove that the attributes more

influencing their overall evaluation also change before and after their reflection.

- To identify if socioeconomic and travel behaviour variables have influence on the overall service quality.

- To compare the importance of the attributes stated by the passengers in the survey, with the importance derived by the models.

- To extract decision rules about the passengers' service quality evaluation that provide public transport managers useful and practical information for formulating strategic transport policies. The attributes were recoded in a 3-point semantic scale in order to achieve more practical rules.

Study Case 2: Decision Trees stratifying the sample in the bus public service. The aims of this study case were:

- To analyze the Overall Evaluation of the service among segments of the market that share some characteristics. As the global sample has to be divided in various subsamples, a large number of data is needed. For this reason, data from various CSSs were aggregated in order to achieve an adequate size of the subsamples at each segment analyzed.

- To discover the main differencies about the service quality evaluation among segments, regarding the structure of the tree, as well as in the most important variables identified at each segment The independent variables were not recoded in a 3-point semantic scale, because the differencies could be hidden with the recodification.

- To compare the importances of the attributes derived by the models by market segment with the importances of the attributes stated by the users in the Ranking scale (only three attributes were highlighted by each passenger)

Study Case 3: Decision Trees for the rail public service. The main objectives of this study case were:

- To validate the use of the methodology Decision Trees for analyzing service quality in other context: other public transport mode and other country.

- To analyze the passengers' Overall Evaluation about the service among different market segments according to common criteria of segmentation, such as the frequency of use or the type of user, but also other criteria less common, such as the type of day or time of the trip.

- To extract useful and practical information for public transport managers.

At the three study cases the data analyzed were collected by non-research oriented surveys. These surveys were developed with the unique purpose of a simple statistical frequency analysis by the transport providers. In addition, the common objetive of the three study cases was to validate the use of Decision Trees for analyzing service quality in public transportation.

5.1. Study Case 1: Decision Trees for the Pre-Evaluation and Post-Evaluation of the bus public service

The main purpose of this doctoral thesis is to examine whether or not the CART model can effectively evaluate service quality in public transportation and also identify the key factors affecting this concept. In this first part of the experimental context for the bus transit service, this is also one of the main research objetives, due to the fact that this is the first time that a CART model has been applied to a bus public transport with service quality analysis purposes. Another aim of this section is to verify the hypothesis of dell'Olio et al (2010) regarding the different evaluation passengers make of service attributes before and after making them reflect on those attributes. In order to achieve these objetives, the data from the CSS conducted in the Granada metropolitan transit system in 2007 were used in this section. This data were collected in a non-research oriented survey, which later, will be demostrated that it can be used in researching critical elements and it provides an approach to increasing the collaboration between researchers and the public transport industry.

Two different models were built to classify the dependent variable (Pre- and Post-Evaluation) and identify the attributes that play a key role in the classification of this variable. Model 1 (Pre-Evaluation CART) allows the identification of the variables which, a priori, are more important in passengers' perception of SQ. Model 2 (Post-Evaluation CART) shows the variables that take importance once the passengers have reflected on the characteristics defining the service. 21 variables were used as independent variables for the models. 3 variables about the customers' demographic profile (gender, age and occupation), 6 about their travel behaviour (reason for travelling, frequency of use, type of ticket, availability of a private vehicle, complementary modes used for access to the bus stop, complementary modes used for moves from the bus stop) and 12 service quality attributes.

The stated importance for each attribute indicated by the respondents was compared with the derived importance obtained from the CART algorithm in the two models (Pre- and Post-Evaluation model).

5.1.1. Data preparation

One of the preprocessing filter used on this dataset (data from the metropolitan bus service in 2007) was to delete the observations in which one of the target variable Pre-Evaluation or Post-Evaluation had missing values. Among the 1,200 total number of observations of the sample, the global dataset was reduced to 858 valid observations.

Moreover, in order to find results that could be easier to interpreter for Public Transport managers, as well as to homogeneize the scales used by both target variables (Pre-Evaluation and Post-Evaluation) and the service quality attributes, the rates were recoded into a reduced semantic scale. It was a 3-point semantic scale comprising POOR, FAIR and GOOD as levels of the service quality. The recodification of the target variable in a more reduced scale was also performed by Huang and Hsueh (2010), who converted the four classes of the target variable into two classes (Bad and Good) in order to find out more applicable association rules.

So, for the Pre-Evaluation variable, the recodification was performed according to: rates 1 and 2 as POOR, 3 as FAIR and 4 and 5 as GOOD. The recodification of the target variable Post-Evaluation as well as the one developed for the service quality attributes used as predictors of the models, was carried out in the same way because they used the same numeric scale (a 11-point likert scale). They were recoded comprising the values 0 to 3 as POOR, from 4 to 6 as FAIR and from 7 to 10 as GOOD.

First, we present a brief analysis of the variables' values recoded to the new scale. Figure 8 shows that the rates assigned to the overall service quality changes among them (Pre-Evaluation and Post-Evaluation). When passengers have reflected on the attributes defining the service, their proportion of passengers evaluating the service as GOOD or POOR diminish, while the proportion of acceptable evaluations (FAIR) increase. Moreover, this figure shows the imbalanced nature of the dataset, being the POOR overall evaluation of the service the subset of data with the lowest number of cases.

Figure 9 shows the distribution of the perceptions rates stated by the users about the SQ attributes used, as independent variables, in both models, recoded into the new semantic scale. Figure 9 shows that a 88.4% of the sample perceives this Information with a calification of Fair or better, and only a 17.6% perceived it as Poor. Punctuality presents also a high valoration among users, with a large number of passengers considering it as Good, and a low number of passengers dissatisfied with its performance (only 11.9% perceived it as Poor). This distribution is similar in most part of the variables, in which the number of users with a Good perception of the variable is, in general, higher than the ones that have a Fair perception, and only a small part of the sample thinks that the performance of this attribute is Poor. However, the distribution of observations in the variable Fare changes from the usual distribution found out in the other variables. In this case, the number of passengers that perceived this variable as Poor suffered a large increased (with a percentage of 22.3% from the total) and the passengers that consider the Fare as Good is also minor than the number expected following the previous disributions of other variables (only a 33.1%).

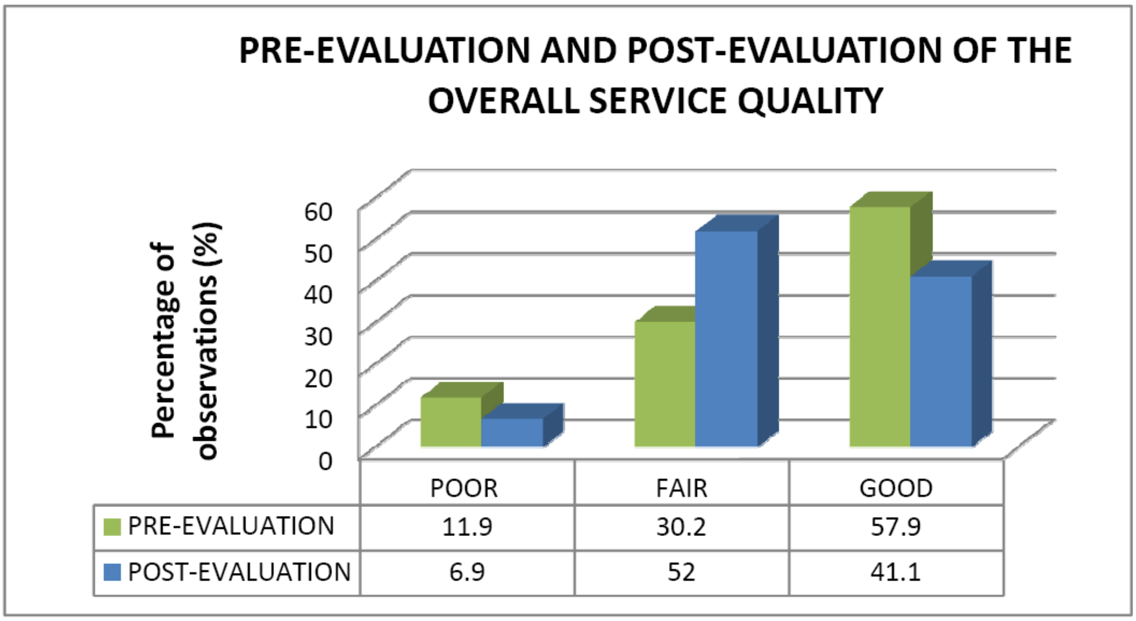
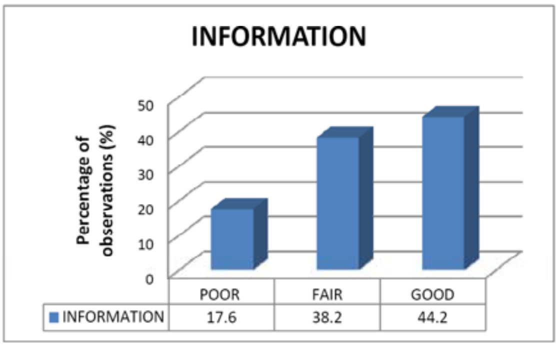
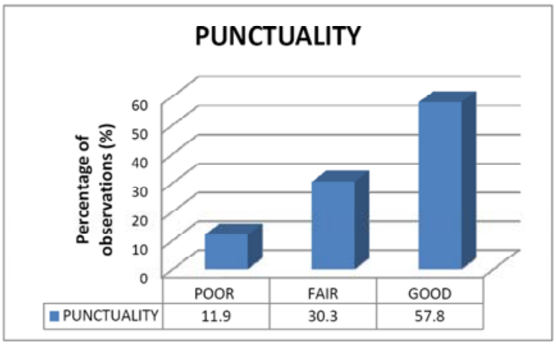


Figure 8. Pre-Evaluation and Post-Evaluation of the overall service quality perceived by passengers

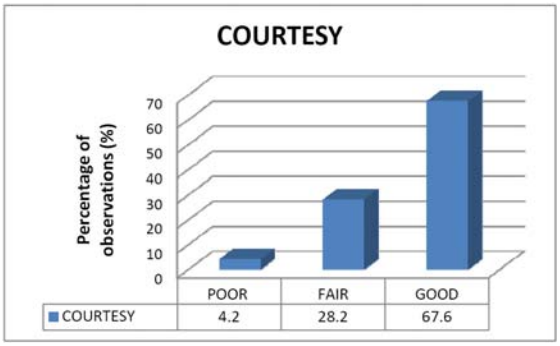
The variables that have the lowest number of SQ observations considered as Poor are Courtesy, Safety and Cleanliness, with values of 4.2%, 6.5% and 6.8% respectively. On the contrary, the attributes that have the highest proportion of observations considered as Poor are Fare (with a 22.3% of the cases), Information (with a 17.6%), Accesibility (16.9%) and Frequency (16.3%).

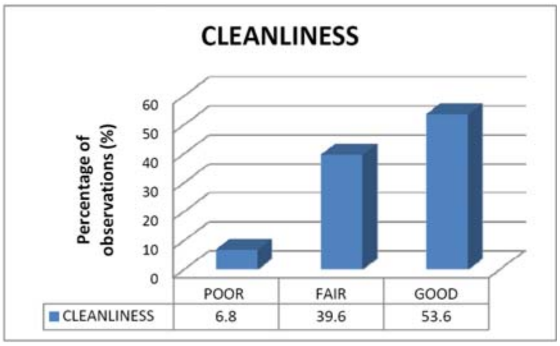
By observing the percentage of Good observations among attributes, Fare, Information and Accesibility are the ones with the lowest proportion (33.1%, 44.2% and 46.2% respectively) coinciding with the ones that obtained also the higher number of observations classified as Poor, then consequently, their average evaluation will be lower than in other attributes. On the other hand, the attributes that have the highest number of Good observations are the Courtesy (67,6%) and the Safety (63,6%), coinciding also with the attributes with the minor number of Poor observations.

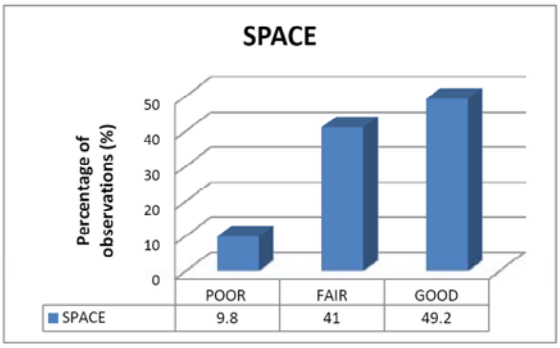

INFORMATION
Percentage of observations (%)
0
10
20
30
40
50
POOR
FAIR
GOOD
INFORMATION
17.6
38.2
44.2

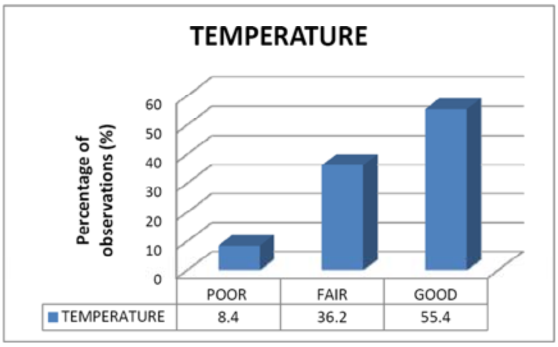

PUNCTUALITY
Percentage of observations (%)
0
10
20
30
40
50
60
POOR
FAIR
GOOD
PUNCTUALITY
11.9
30.3
57.8

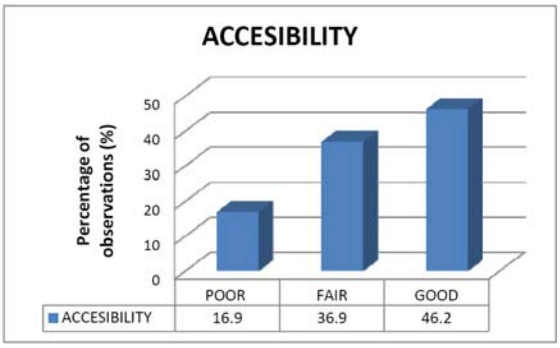

SAFETY
Percentage of observations (%)
0
10
20
30
40
50
60
70
POOR
FAIR
GOOD
SAFETY
6.5
29.8
63.6

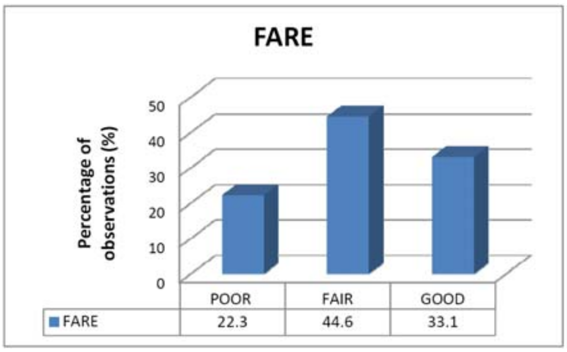

COURTESY
Percentage of observations (%)
0
10
20
30
40
50
60
70
POOR
FAIR
GOOD
COURTESY
4.2
28.2
67.6

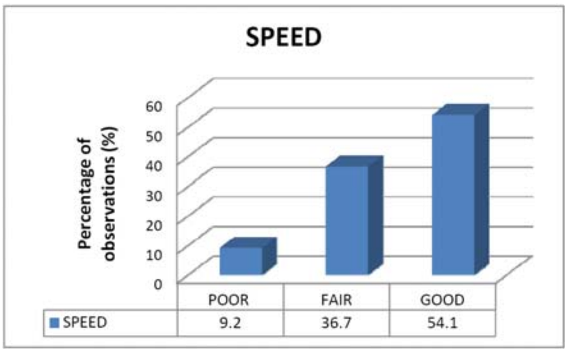

CLEANLINESS
Percentage of observations (%)
0
10
20
30
40
50
60
POOR
FAIR
GOOD
CLEANLINESS
6.8
39.6
53.6


SPACE
Percentage of observations (%)
0
10
20
30
40
50
POOR
FAIR
GOOD
SPACE
9.8
41
49.2


TEMPERATURE
Percentage of observations (%)
0
10
20
30
40
50
60
POOR
FAIR
GOOD
TEMPERATURE
8.4
36.2
55.4


ACCESIBILITY
Percentage of observations (%)
0
10
20
30
40
50
POOR
FAIR
GOOD
ACCESIBILITY
16.9
36.9
46.2


FARE
Percentage of observations (%)
0
10
20
30
40
50
POOR
FAIR
GOOD
FARE
22.3
44.6
33.1


SPEED
Percentage of observations (%)
0
10
20
30
40
50
60
POOR
FAIR
GOOD
SPEED
9.2
36.7
54.1

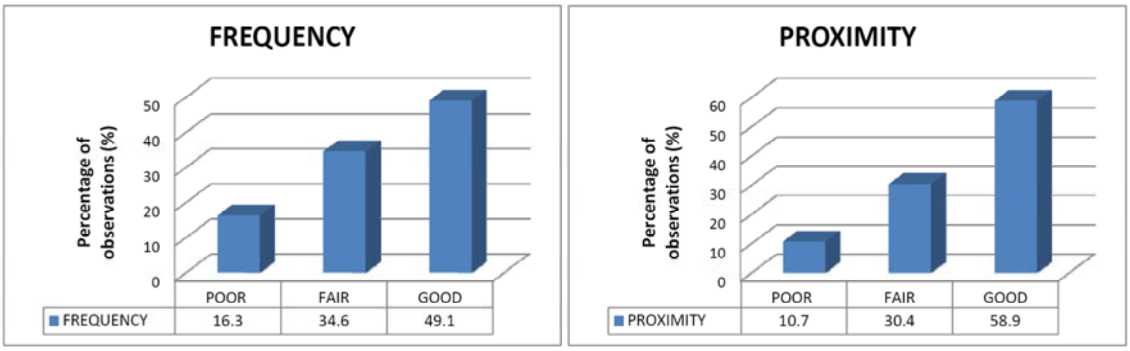


Figure 9. Recodification of the attributes

5.1.2. Decision Tree

Figure 10 shows the CART model for the dependent overall Pre-Evaluation SQ variable. The interpretation of the tree is given below. A root node (Node 0) is divided into two child nodes (Node 1 and Node 2). It is used as a splitter the variable that obtains the maximum 'purity' of the two child nodes. In this case, the splitter is Frequency. Node 1 shows the data related to passengers who have a Good or Fair perception of service Frequency. In turn, Node 1 is divided into two terminal nodes or child nodes (Node 3 and Node 4) on the basis of the Punctuality variable. Terminal Node 3 shows that if Punctuality and Frequency are rated as Good or Fair, the overall evaluation of SQ (Pre-Evaluation) is likely to be perceived as GOOD (67.8%). Terminal Node 4 shows that if Frequency of service is rated as Good or Fair and Punctuality is stated as Poor, there is a 45.8% likelihood that the occupant will consider that the global SQ is FAIR.

The passengers who have a Poor perception of service Frequency are on the right branch of the tree. In this case, Node 2 is divided into two terminal nodes (Node 5 and Node 6). Terminal Node 5 indicates that a passenger who travels for a reason other than Occupation, Studies or Doctor, and who rates service Frequency as Poor, will rate SQ as FAIR in 50.8% of cases. If the reason for travelling is Occupation, Studies or Doctor (i.e. compulsory mobility) and Frequency has been rated as Poor (Terminal Node 6), the evaluation of SQ will be POOR in 49.4% of cases.

This first decision tree produced two levels (depth below the root node), 7 nodes and 4 terminal nodes or leaves. A 10-fold cross-validation of the sample was used to give us an accuracy indicator of the categorization of the variable class of 59.72%, indicating that the model's precision was acceptable (Wong and Chung, 2007)

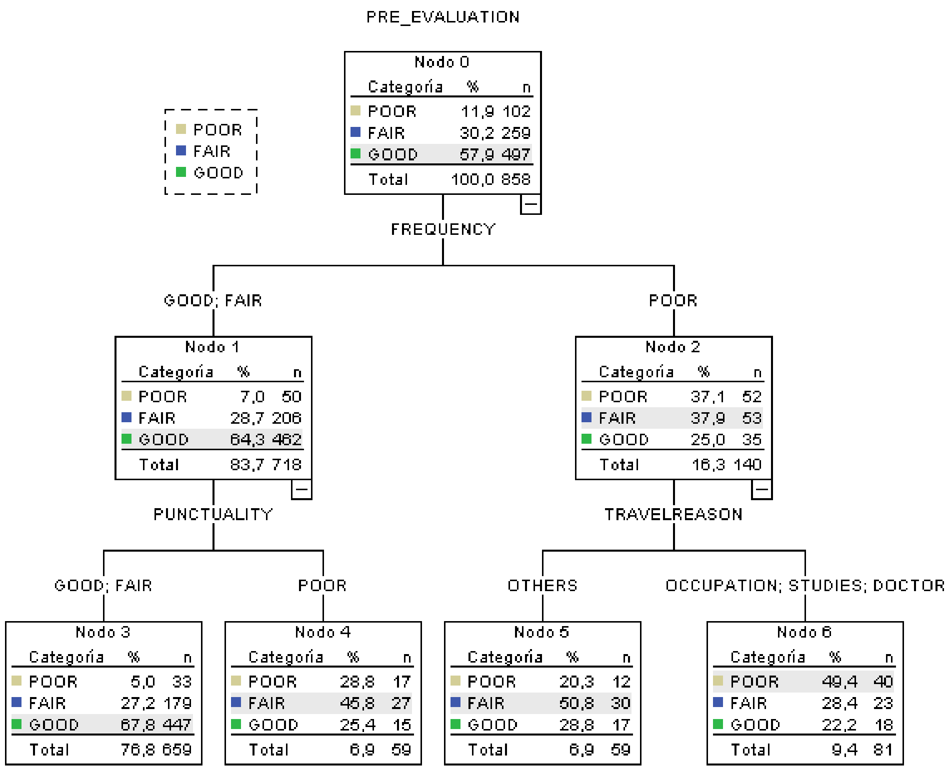


Figure 10. Pre-evaluation CART

The CART built for the overall Post-Evaluation SQ variable produced 5 levels, 23 nodes and 12 terminal nodes (see Figure 11). In this case, the root node divides into 2 child nodes after the Punctuality variable. The data of the passengers who have a Good perception of service Punctuality are on the left branch of the tree, giving 6 terminal nodes (8, 10, 15, 16, 17 and 18). All these terminal nodes predict that passengers will rate SQ as GOOD or FAIR. This implies that a passenger who rates service Punctuality as Good will give the service a (Post-Evaluation) global evaluation of FAIR or higher. If the same passengers who rated Punctuality as Good also rate Proximity and Safety as Good and Fare is not rated as Poor (Terminal Node 15), it is very likely that the overall evaluation of the service will be GOOD (76.2%). On the other hand, if Punctuality, Speed and Fare are rated as Good, the overall evaluation of the service is likely to be GOOD (Terminal Node 17, 67.9%), even if Proximity is considered Poor or Fair.

The passengers who have a Poor or Fair impression of the Punctuality variable are on the right branch of the tree, where 6 terminal nodes are obtained (12, 13, 14, 20, 21 and 22). On the basis of the Frequency variable, Node 2 is divided into 2 child nodes: Node 5 for the passengers who rate Frequency as Good or Fair and Node 6 for all other passengers.

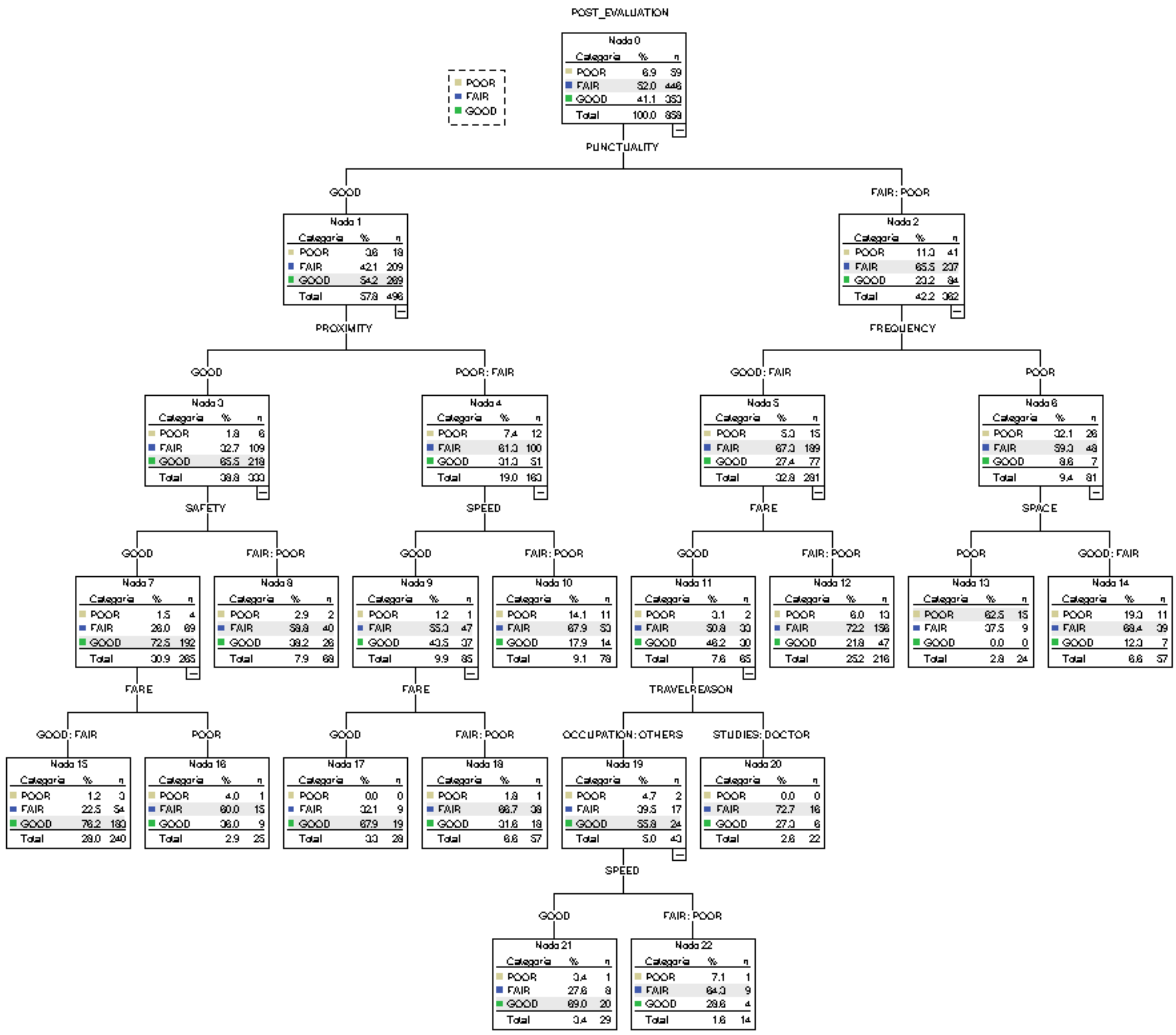


Figure 11. Post-evaluation CART

If Punctuality is not Good and Frequency is rated as Poor, passengers will probably not give the service a GOOD overall evaluation. After Node 6, Space is used as a splitter and 2 terminal nodes are obtained (Nodes 13 and 14). Terminal Node 13 shows that if Punctuality is not Good and Frequency and Space are Poor, the overall evaluation of the service will probably be POOR (62.5%). In the event that Space is not Poor, the global service evaluation will probably be FAIR (68.4%, Terminal Node 14).

After Node 5 (passengers who do not rate Punctuality as Good but do not consider Frequency to be Poor) four terminal nodes are obtained (12, 20, 21 and 22), in which the overall evaluation of the service is FAIR or higher. This implies that if the service provided gives passengers the impression that Frequency is Fair or Good, their global evaluation of the service will not be POOR, even if Punctuality is not Good. Moreover, if the passengers perceive service Fare and Speed as Good and the Reason to travel is Occupation or Other, the overall evaluation of the service will probably be GOOD (69.0%, Terminal Node 21).

The Post-Evaluation CART shows that the model has a global accuracy indicator value of 62.16%, which indicates that the model is stable and its precision is acceptable. The precision of this model is somewhat higher than the precision of the previous model, which implies that overall Post-Evaluation SQ can be predicted more accurately than overall Pre-Evaluation SQ.

5.1.3. Decision rules

One of the main advantages of decision trees, as opposed to other modelling methods, is that they provide effective "If-then" rules that make the model very practical and easy to interpret from the perspective of management by PT operators and managers.

Each decision tree gives as many rules as the existing number of terminal nodes. Table 13 shows the 4 rules from Decision Tree One (Pre-Evaluation CART), which uses the variables Frequency, Travel reason and Punctuality. One of the rules identifies the conditions that must be given for the overall evaluation of service to have a high likelihood of being considered GOOD (Node 6). In this model, two rules for an overall evaluation of FAIR and one rule for an evaluation of POOR were identified. The confidence rate of the rules is not very high (about 50% for the Nodes 4, 5 and 6), and only in the Node 3 is reached a good value (67.8%). The support rate obtained an acceptable representation of the sample in all the rules (values higher than 3%), higher than the threshold value established in others studies, such as Montella et al. (2012) who used a threshold rate of 0.10% or de Oña et al. (2013) who used a threshold rate of 0.60%.

Table 13. Rules for overall Pre-Evaluation of service quality

NODE	RULE		CONFIDENCE RATE (%)	SUPPORT RATE (%)
	IF	THEN		
3	“Frequency” and “Punctuality” are rated as Good or Fair	Service is rated as "Good"	67.8	52.1
4	"Frequency" is rated as Good or Fair and the perception of “Punctuality” is Poor	Service is rated as "Fair"	45.8	3.1
5	"Frequency" is rated as Poor, and the "Travel Reason" is Others	Service is rated as "Fair"	50.8	3.5
6	"Frequency" is rated as Poor, and the "Travel Reason" is different to Others	Service is rated as "Poor"	49.4	4.7

Table 14 shows the 12 rules of Decision Tree Two (Post-Evaluation CART) that use the attributes Punctuality, Frequency, Proximity, Space, Fare, Speed, Safety and

Travel reason to identify rules that are useful to service managers. It bears mentioning that only one rule was found to imply a high probability that the overall evaluation of service will be POOR (Node 7): If Frequency and Space are perceived as Poor and Punctuality is not Good, the overall evaluation of service is likely to be POOR (62.5%). On the contrary, three rules for GOOD evaluations and eight rules for FAIR evaluations were identified. Finally, it can be seen that the confidence values of the rules taken from the Post-Evaluation CART are higher than the ones taken from the Pre-Evaluation CART, with a minimum value of 58.8%. By observing the support rates, the minimun values is obtained in the Node 22 with only including a 1% of the sample.

Table 14. Rules for overall Post-Evaluation of service quality

NODE	RULE		CONFIDENCE RATE (%)	SUPPORT RATE (%)
	IF	THEN		
15	"Punctuality" is rated as Good, "Proximity" and "Safety" are rated as Good, and there is a perception of "Fare" as Good or Fair	Service is rated as "Good"	76.2	21.0
17	"Punctuality" is rated as Good, "Proximity" is rated as Poor or Fair, and there is a perception of “Speed” and "Fare" as Good	Service is rated as "Good"	67.9	2.2
21	"Punctuality" is rated as Poor or Fair, "Frequency" is not rated as Poor, the perception of "Fare" is Good, the "Travel Reason" is Occupation or Others, and "Speed" is rated as Good.	Service is rated as "Good"	69.0	2.3
8	"Punctuality" and "Proximity" are rated as Good but the perception of "Safety" is other than Good.	Service is rated as "Fair"	58.8	4.7
10	"Punctuality" is rated as Good, and "Proximity" and "Speed" are rated as Poor of Fair.	Service is rated as "Fair"	67.9	6.2
12	"Punctuality" is rated as Poor or Fair, "Frequency" is not rated as Poor, and the perception of "Fare" is Poor or Fair	Service is rated as "Fair"	72.2	18.2
14	"Punctuality" is rated as Poor or Fair, "Frequency" is rated as Poor, and the perception of "Space" is Good or Fair	Service is rated as "Fair"	68.4	4.5
16	"Punctuality", "Proximity" and "Safety" are rated as Good, and there is a perception of "Fare" as	Service is rated as "Fair"	60.0	1.7

	Poor			
18	"Punctuality" is rated as Good, "Proximity" is rated as Poor or Fair, there is a perception of "Speed" as Good, and "Fare" is rated as Poor or Fair	Service is rated as "Fair"	66.7	4.4
20	"Punctuality" is rated as Poor or Fair, "Frequency" is not rated as Poor, the perception of "Fare" is Good, and the "Travel Reason" is Studies or Doctor	Service is rated as "Fair"	72.7	1.9
22	"Punctuality" is rated as Poor or Fair, "Frequency" is rated as other than Poor, the perception of "Fare" is Good, the "Travel Reason" is Occupation or Others, and "Speed" is rated as Poor or Fair	Service is rated as "Fair"	64.3	1.0
13	"Punctuality" is rated as Poor or Fair, "Frequency" is rated as Poor, and the perception of "Space" is Poor	Service is rated as "Poor"	62.5	1.7

5.1.4. Importance of the variables

The CART modelling process has a crucial phase in which the variables that are of key importance in the prediction of the dependent variable are identified. This is achieved by using the importance index (Kashani and Mohaymany, 2011), of which a standardized form is used in this research work to reflect the importance of each independent variable on the model.

Table 15 shows the standardized importance of the attributes deduced from each of the two models (Pre- and Post-Evaluation CARTs), and the importance stated by passengers in the CSS. It can be seen that there is very little variation in the evaluations stated by the passengers in the surveys, considering that all the attributes are highly important. The average value of attribute importance in the CSS is concentrated in the 8.5 to 9.5 range (on a scale of 0 to 10). Therefore, their standardized 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 (Weinstein, 2000).

Analysing the importances derived from the Pre-Evaluation CART, Frequency is the attribute with the highest weight, far from the other attributes. A priori, this would imply that passengers rate SQ on Frequency alone. Eboli and Mazzulla (2008b; 2010) also identified service Frequency as the attribute that had the greatest impact on SQ. Dell'Olio et al. (2010; 2011) identified Frequency as one of the most important attributes. The Pre-Evaluation CART reveals that Frequency is a key attribute of major impact when passengers have a preliminary idea of how the service operates. Moreover, Frequency serves as the tree's root variable, splitting the passengers that evaluate SQ as GOOD onto the left branch and the passengers that evaluate it as POOR onto the right branch, while the passengers who give a FAIR evaluation are split in either direction. Speed and Punctuality are also attributes that carry considerable weight on SQ in the Pre-Evaluation CART, although at quite a distance from Frequency. This matches the results of other recent studies (Dell'Olio et al., 2010; Eboli and Mazzulla, 2010) in which Punctuality and service Reliability have been identified as one of the most important attributes for passengers.

After the passengers have reflected on the service attributes, however (because they have been asked about them), a higher number of attributes gain weight in the overall perception of quality, whereas the weight of Frequency on the overall evaluation decreases. Table 15 shows that, apart from the attributes considered to be important in the Pre-Evaluation (Frequency, Speed and Punctuality), other attributes such as Proximity, Safety and Fare are also identified as important in the Post-Evaluation CART when the passengers have been made to reflect on them. They can attain standardized importance values that exceed 60.3%.

Table 15. Stated and derived attributes' importance

STATED IMPORTANCE		DERIVED IMPORTANCE USING CART			
		PRE-EVALUATION		POST-EVALUATION	
PUNCTUALITY	100%	FREQUENCY	100%	PROXIMITY	100%
FREQUENCY	98.9%	SPEED	55.2%	SPEED	78.8%
SAFETY	98.3%	PUNCTUALITY	49.3%	SAFETY	72.3%
CLEANLINESS	96.9%	TRAVELREASON	15.9%	FREQUENCY	68.4%
ACCESSIBILITY	96.8%	USEFREQUENCY	11.6%	FARE	64.1%
COURTESY	95.7%	TICKET	6.4%	PUNCTUALITY	60.3%
FARE	95.5%	PROXIMITY	5.4%	SPACE	46.7%
SPEED	95.5%	AGE	3.5%	COURTESY	42.0%
TEMPERATURE	95.4%	OCCUPATION	0.2%	TEMPERATURE	38.5%
PROXIMITY	95.0%	SAFETY	0	INFORMATION	28.3%
SPACE	94.7%	CLEANLINESS	0	TRAVELREASON	20.4%
INFORMATION	94.1%	ACCESSIBILITY	0	ACCESSIBILITY	17.6%
		COURTESY	0	CLEANLINESS	11.9%
		FARE	0	SEX	5.4%
		TEMPERATURE	0	OCCUPATION	4.6%
		SPACE	0	USEFRECUENCY	2.2%
		INFORMATION	0	MODESFROM	1.6%
		SEX	0	AGE	0.9%
		MODESFROM	0	TICKET	0.8%
		PRIVATEVEHICLE	0	PRIVATEVEHICLE	0.7%

Therefore, after making the passengers reflect on the variables that can have an impact on their perception of the evaluation of PT, the importance of Frequency diminishes and the role of other service attributes in the overall perception of SQ increases, such as Proximity, Safety and Fare. These outcomes match those obtained by Dell'Olio et al. (2010). They compared an overall evaluation of SQ before and after making passengers reflect on the importance of certain fundamental system variables which they may not have previously considered. Ordered probit models were used. In their first model (pre-evaluation), Dell'Olio et al identified Reliability of Service (RS) and Waiting Time (WT) (which could be considered equivalent to Punctuality and Frequency in this research work) as the two variables that had the greatest impact on passenger's overall evaluation of SQ. Likewise, in their second model (post-evaluation), the importance of Frequency diminished as the importance of other attributes increased.

5.2. Study Case 2: Decision Trees stratifying the sample in the bus public service

The main purpose of the research work carry out in this section is to examine whether the evaluation of SQ, as well as the key drivers towards SQ, are different among segment of passengers which share some characteristics.

The research work herein uses the data gathered in four CSS conducted in the Granada metropolitan transit system from 2008 to 2011, which were non-research oriented surveys. The research results bring practical value to the public transport industry by identifying the key factors in each market segment that planners can focus on in their efforts to enhance quality.

Fourteen different CART decision trees were built – one for each market segment – and another CART was built for the overall market. In each one of the CART all the service attributes (12) were considered in the models as well as the variables defining socioeconomic characteristics and travel habits of passengers that had not been used for segmentation (eight variables for the overall market's model and seven variables for the other 14 models).

5.2.1. Data preparation

From the original dataset consisting of data from four CSSs, the observations that presented a missing value on the target variable (overall SQ) or in the socioeconomic characteristics and travel habits variables used for the segmentation, were removed from the whole sample. A total of 3,664 observations were valid for this analysis.

To find results that could be easiest for PT managers to interpret, and taking into account the low frequency of categories 1 (2.5%), 2 (5.9%) and 5 (4.0%), the target variable (overall SQ) was recorded in a reduced semantic scale. It was a three-point semantic scale comprising rates 1 and 2 as POOR, 3 as FAIR, and 4 and 5 as GOOD. In this case, the SQ attributes included in the analysis as predictor of the model used their original 11-point scale.

The following figures (Figure 12 to 17) display the distribution of the Overall SQ perceptions recoded into the new three point semantic scale. As happened in the previous study case, it can be observed that the dataset composing the sample is imbalanced, showing classes that are under represented and classes over represented.

These figures show the percentage of observations about the target variable"Overall Evaluation" that are classified as POOR, FAIR or GOOD across the different market segments under study and also for the overall market. As can be observed, their distribution is very similar across segments, with little proportion of observations considered as POOR, a high number of observations as FAIR, and the most part of the cases classified as GOOD. The market segments that achieved the lowest number of observations classified as POOR were the passegers using the Senior Citizen Pass, with a 4.4% of the cases, and the Male with a percentage of 5.5%. On the other hand, the segments that achieved the highest proportion of observations considered as POOR were the passengers using Other type of ticket (16.7%). By observing the percentage of GOOD observations, its value was good and higher that 60% in all the market segments and also in the overall market. It ranged from a 62.8% for passengers travelling for Studies purposes, to a 79.9% for passengers using a Senior Citizen Pass. In fact, if it is calculated the average value of these Overall Evaluations (considereing POOR as 1, FAIR as 2 ans GOOD as 3), the users that use the Senior Citizen Pass in their trips by bus are the ones that achieve the highest average rate, with a value of 2.76. On the contrary, the group of users that obtain the lowest average rate of the target variable (2.48) are the passegers using Other type of ticket.

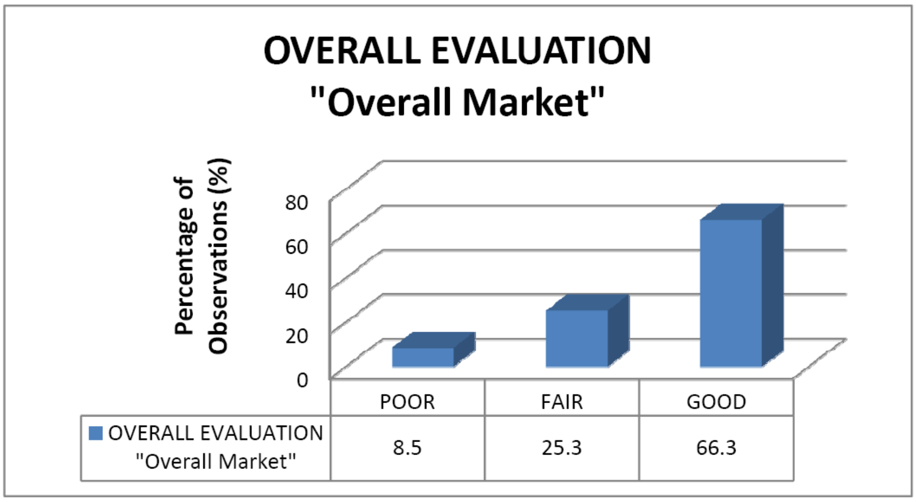


Figure 12. Overall Evaluation for the Overall Market

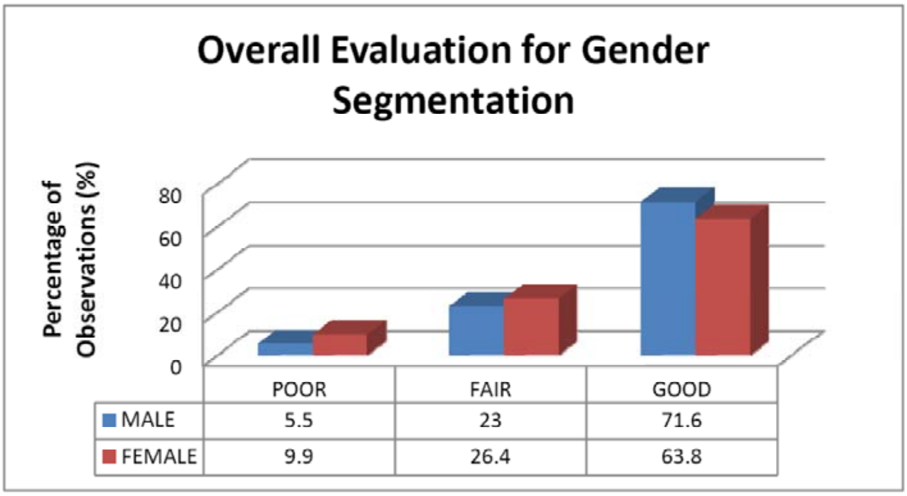


Figure 13. Overall Evaluation according to Gender Segmentation

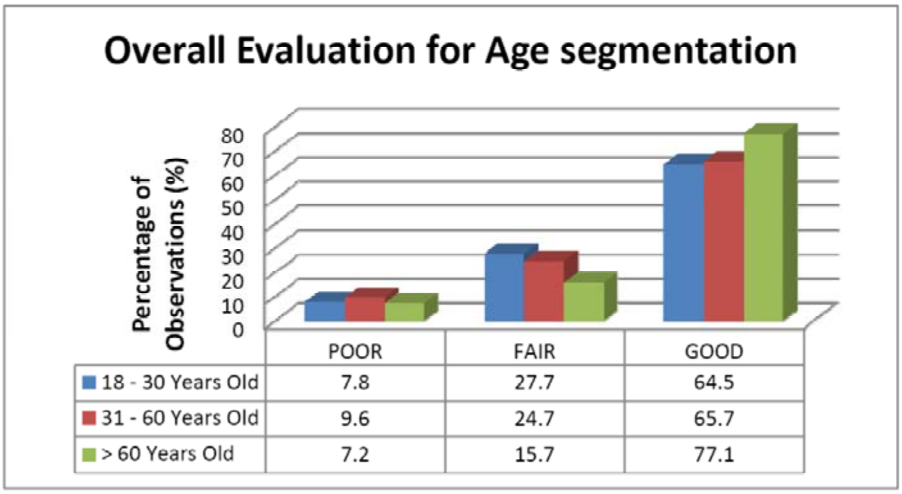


Figure 14. Overall Evaluation according to Age Segmentation

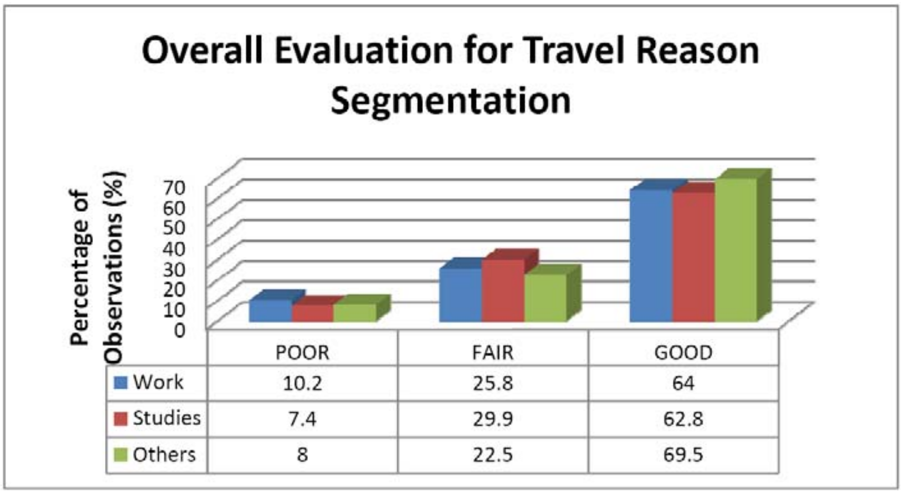


Figure 15. Overall Evaluation according to Travel Reason Segmentation

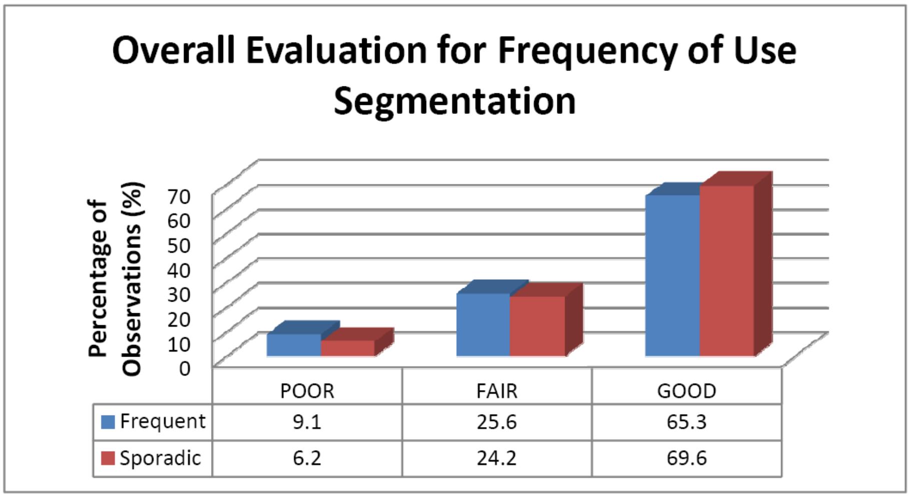


Figure 16. Overall Evaluation according to Frequency of Use Segmentation

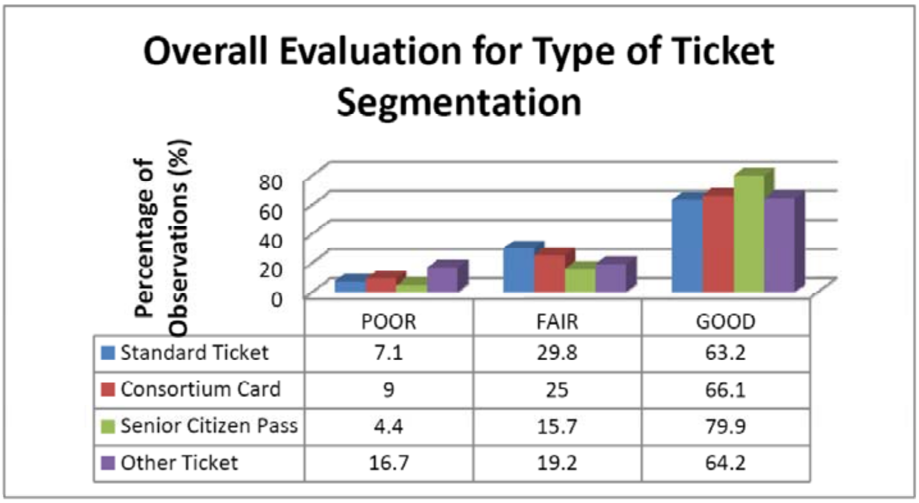


Figure 17. Overall Evaluation according to Type of Ticket Segmentation

5.2.2. Decision Tree

All CART models were built using a 10-fold cross-validation of the sample, which gave an accuracy ratio ranging from 63.65% for Standard Ticket, up to 79.12% for Senior Citizen Pass. The accuracy rates are acceptable for all CART models, and they are higher than the values obtained in other studies in which decision trees were applied for SQ analysis (de Oña et al., 2012; Wong and Chung, 2007).

5.2.2.1. CART for the overall market

Figure 18 shows the CART for the overall market. The root node (Node 0) is split into two child nodes (Node 1 and Node 2), using the variable that maximizes 'purity' in the two child nodes. In this case, the splitter was Information. When Information is rated with a score higher than 6 (Node 2), the overall SQ is likely to be perceived as GOOD (75.7%). 72.1% of the sample is concentrated in this child node (Node 2), which demonstrates that this factor is a great discriminant of the model. The next best splitting criterion for those who scored Information with a value equal to or lower than 6 is Frequency. This is a key variable for discriminating user perception of overall SQ. It groups those who give a value of the Overall Evaluation of POOR or FAIR on the left side (Nodes 5 and 6), as opposed to those who rate it as GOOD or FAIR, on the right side (Nodes 8, 9 and 10). The cut-off point for Frequency is a value of 2. When perceived Frequency is very bad (≤2) and Proximity is considered insufficient (≤4), there is a high probability (69.7%) that the passenger will rate SQ as POOR. On the other hand, if the Frequency scores higher than 2 and Temperature has an adequate score (>6), SQ perception will be GOOD. When Frequency scores high enough (>6), a rating of GOOD is obtained even when the score for Temperature is 6 or lower. This tree is 68.56% accurate.

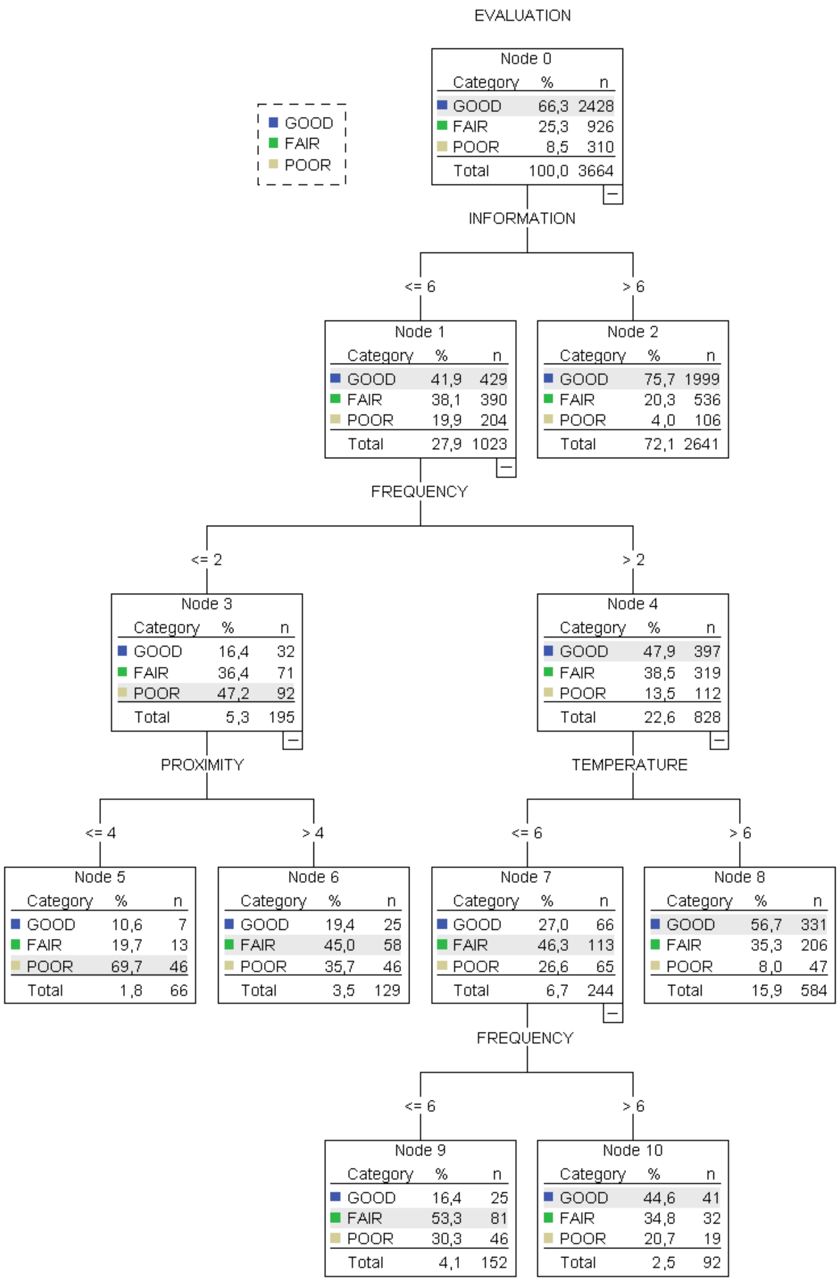


Figure 18. CART for the metropolitan public transport in Granada (Spain). Overall market (Data from 2008 to 2011)

5.2.2.2. CART for the different market segments

14 market segments (Gender, Age, Use Frequency, Travel Reason and Type of Ticket) based on the available data (see Table 16) are analyzed. A CART was built specifically for each one of the segments. Section 5.2.2.2.1 shows the CART models for Male and Female. Section 5.2.2.2.2 shows the CART models for Young people (interval age in {18-30}), Middle age ({31-60}) and Elderly people ({>60}). Section 5.2.2.2.3 shows the CART models for Frequent and Sporadic passengers (separated at 1 trip per week). Section 5.2.2.2.4 shows the CART models for people who take the bus for Working, Studying and Other Reasons. And section 5.2.2.2.5 shows the CART models for different types of tickets used by the passengers (Standard Ticket, Consortium Pass, Senior Citizen Pass and Other Tickets).

Table 16. Size of each group of users (Bus public service, data from 2008-2011)

Criteria of Classification	*Category of User*	*Number of Observations*	*% of the Sample*
1. Gender	Male	1,171	32%
	Female	2,493	68%
2. Age	Young	1,809	49.4%
	Middle	1,479	40.4%
	Elderly	376	10.3%
3. Frequency of use	Frequent	4,542	78.3%
	Sporadic	1,064	21.7%
4. Travel reason	Working	1,027	28%
	Studying	911	24.9%
	Other reasons	1,726	47.1%
5. Type of ticket	Standard ticket	850	23.2%
	Consortium Pass	2,445	66.7%
	Senior Citizen Pass	249	6.8%
	Other Ticket	120	3.3%

5.2.2.2.1. Gender market segments.

As Figure 19 and Figure 20 show, the splitter variable that best splits the root node for men and women is different. In the model for men the variable is Information. As occurred in the global model (see Figure 18), when the variable obtains a good value (>6) the overall SQ perception is GOOD (80.5%). This model uses Temperature, Safety, Speed and Fare as successive splitter variables. It bears mentioning that when the score for Information and Temperature is equal to or under 6, overall SQ is likely to be perceived as POOR or FAIR (Nodes 5, 7 and 9). However, when Speed and Fare have an acceptable score (>4), the service perception is GOOD (Node 10), providing that Safety scores higher than 2. This model is 71.98% accurate, which is higher than the model generated for the entire model.

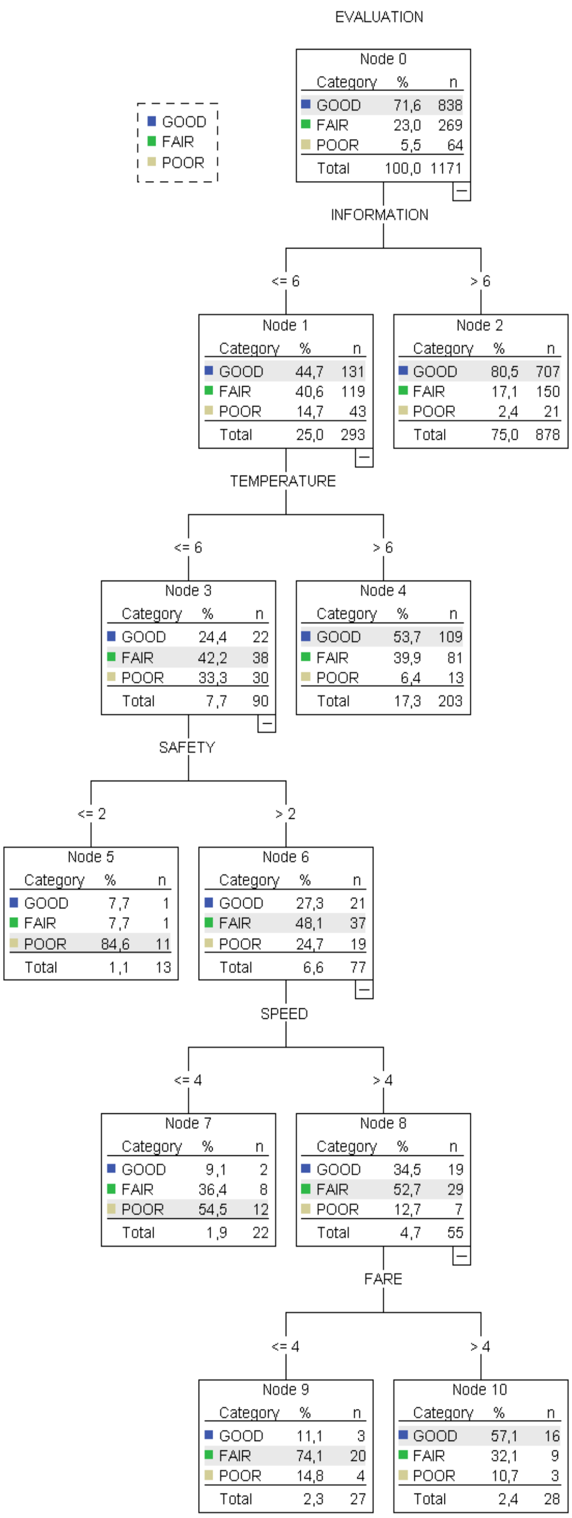


Figure 19. CART for users classified according to the gender (Male)

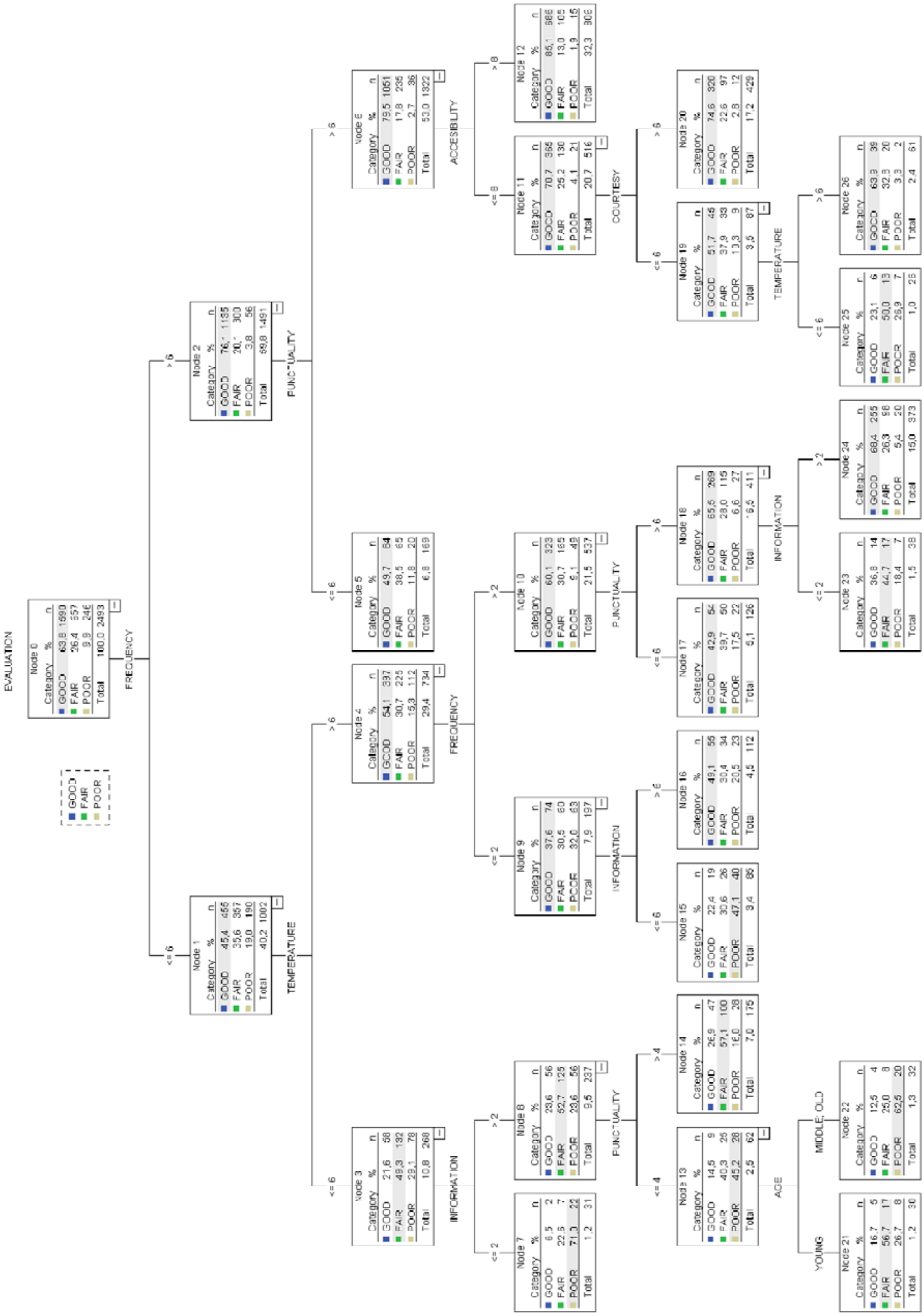


Figure 20. CART for users classified according to the gender (Female)

If we focus on the model developed for women (65.26% accuracy), Frequency splits the tree into two branches. On the right side of the tree, where Frequency has been rated positively (>6), all the leaf nodes obtain *GOOD* values of the variable class, with the exception of Node 25, in which Accesibility, Courtesy and Temperature condition the category selected to a *FAIR* value. On the left side of the tree, Temperature is the attribute that discriminates SQ the best. For values lower than or equal to 6, the variable class obtains a value of *POOR* or *FAIR*, whereas for Temperature values higher than 6, the variable class will be mainly *GOOD* or *FAIR*, even if Frequency is very poorly rated (≤2). The only exception to this is in Node 15, where a *POOR* overall SQ is predicted. The Frequency in this node is very poor (≤2) and Information is insufficient (≤6).

5.2.2.2.2. Age market segments.

Figures 21 to 23 show the models for age-related market segments.

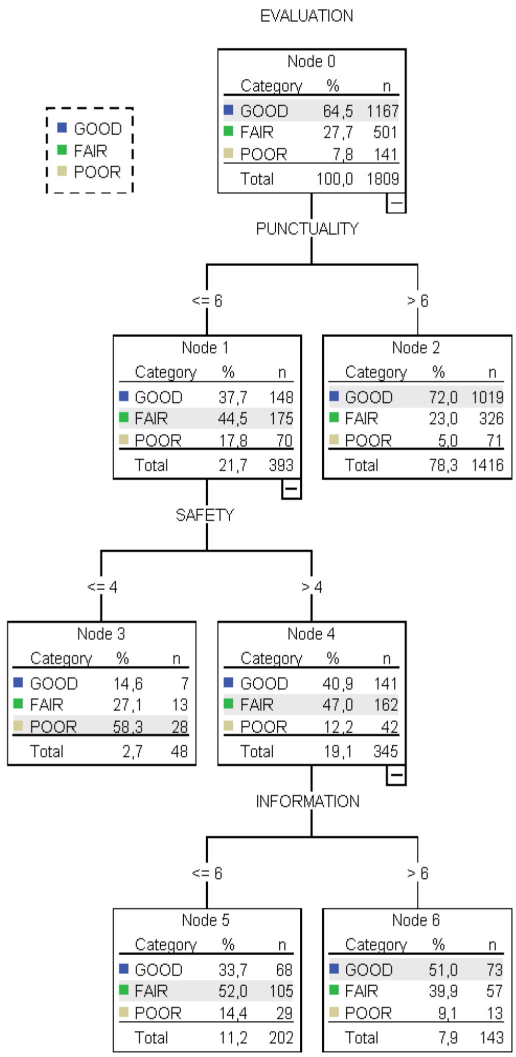


Figure 21. CART for users classified according to the age (Young)

In the tree generated for the Young People, Punctuality is the variable splitter that best discriminates perceived quality, compared to Frequency for Middle-aged and Information for Elderly. This may be because most of the individuals under age 30 (50.4%) are students with not very adaptable schedules, who expect the service to be on time. Individuals of age 30-60 have more flexible schedules, so they attach more importance to Frequency. For the elderly (>60 years old), most of them retired, Punctuality and Frequency are less important, whereas they focus more on good information on the service. It is worth pointing out that Node 7, the model for the elderly, is a pure Node in which quality is rated as POOR in all cases. This occurs after a series of evaluations on Information and Speed that end in the key factor of Proximity which, if it is not rated as good (≤6), the global evaluation of quality will not be good either. The accuracy obtained in these models is 64.73% for Young people, 68.15% for Middle-aged and 78.98% for Elderly.

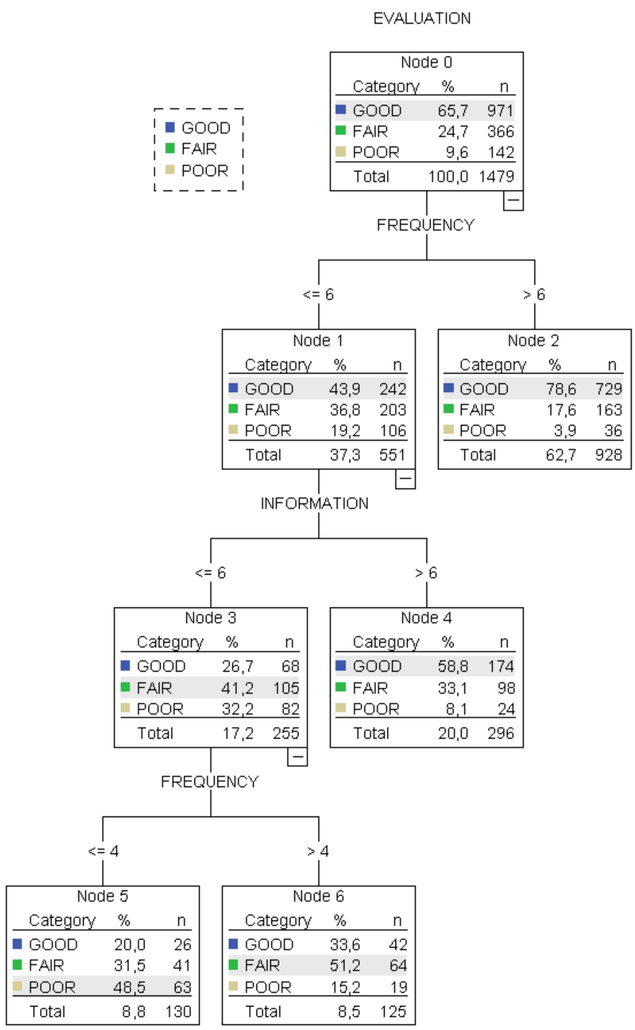


Figure 22. CART for users classified according to the age (Middle)

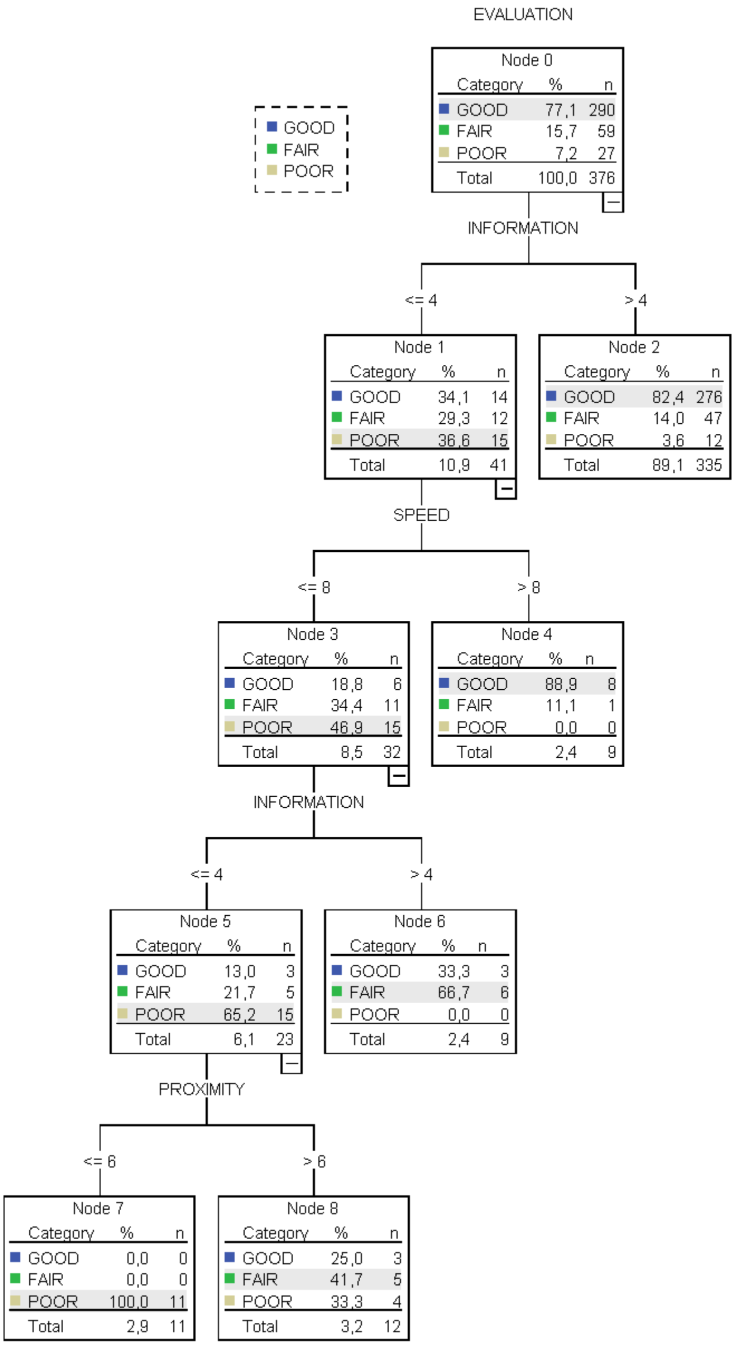


Figure 23. CART for users classified according to the age (Old)

5.2.2.2.3. Frequency of use market segments.

The results in Figure 24 and Figure 25 show that service quality for frequent travelers may be explained by the model created for the overall market, after pruning a few of its branches. This may be because most of the passengers interviewed (78.3%) use the service constantly, and make up a large percentage of the overall sample. It would be erroneous to think that frequent users are not worried about the quality of the Information because they know how to use the system well because of their repeated trips. Quite to the contrary, any sudden changes in itineraries and time tables are more suffered by them than by other users and therefore the Information factor is decisive in their assessment of quality.

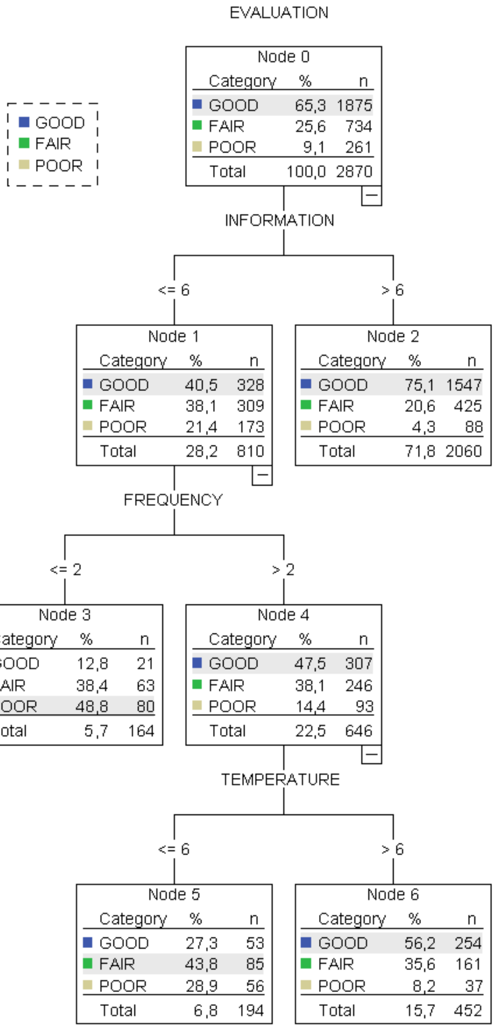


Figure 24. CART for users classified according to the frequency of use (Frequent passengers)

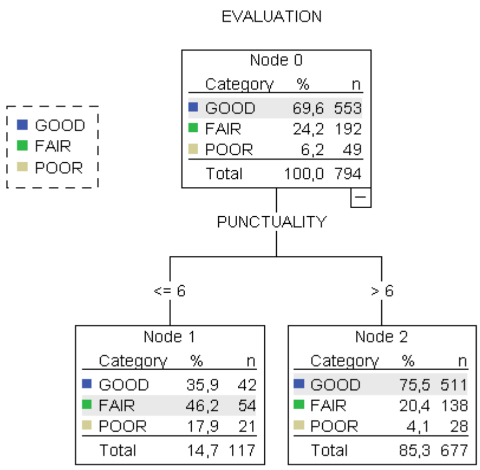


Figure 25. CART for users classified according to the frequency of use (Sporadic passengers)

Quality for sporadic passengers, however, may be best explained in terms of Punctuality. When users take a bus occasionally, they are only concerned with the bus being on time, and pay less attention to other features. The accuracy of the models for frequent and sporadic passengers is 68.29% and 69.77%, respectively.

5.2.2.2.4. Travel reason market segments.

The models determined for the Travel Reason (Figures 26 to 28) underscore what was interpreted in trees for different age groups. This is because when we refer to Young people {18-30}, the segment encompasses most of the segment that has Studies as the Travel Reason, and when we target the population of age 30-60, most of them are travelling for Working reasons. When the reason to travel is Studies, the most important variable is Punctuality. When the reason to travel is Working, Frequency becomes the most discriminant variable. When passengers travel for reasons other than the two preceding ones (Others), Information becomes the most important attribute. These models are 65.86%, 67.67% and 69.58% accurate, respectively.

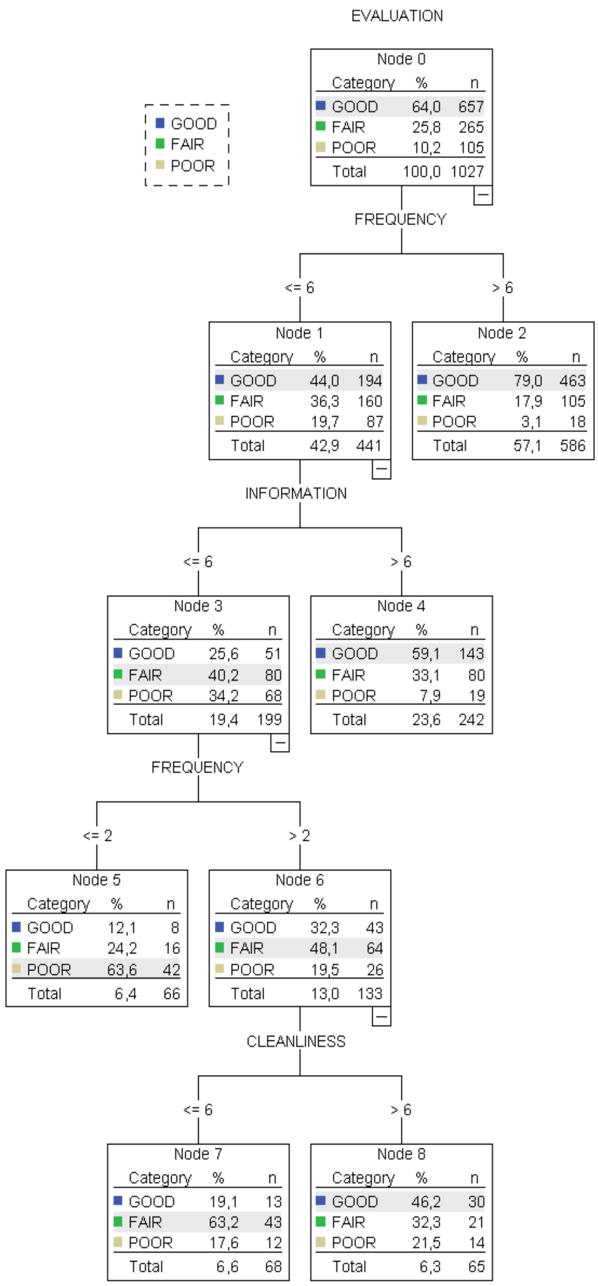


Figure 26. CART for users classified according to the travel reason (Working)

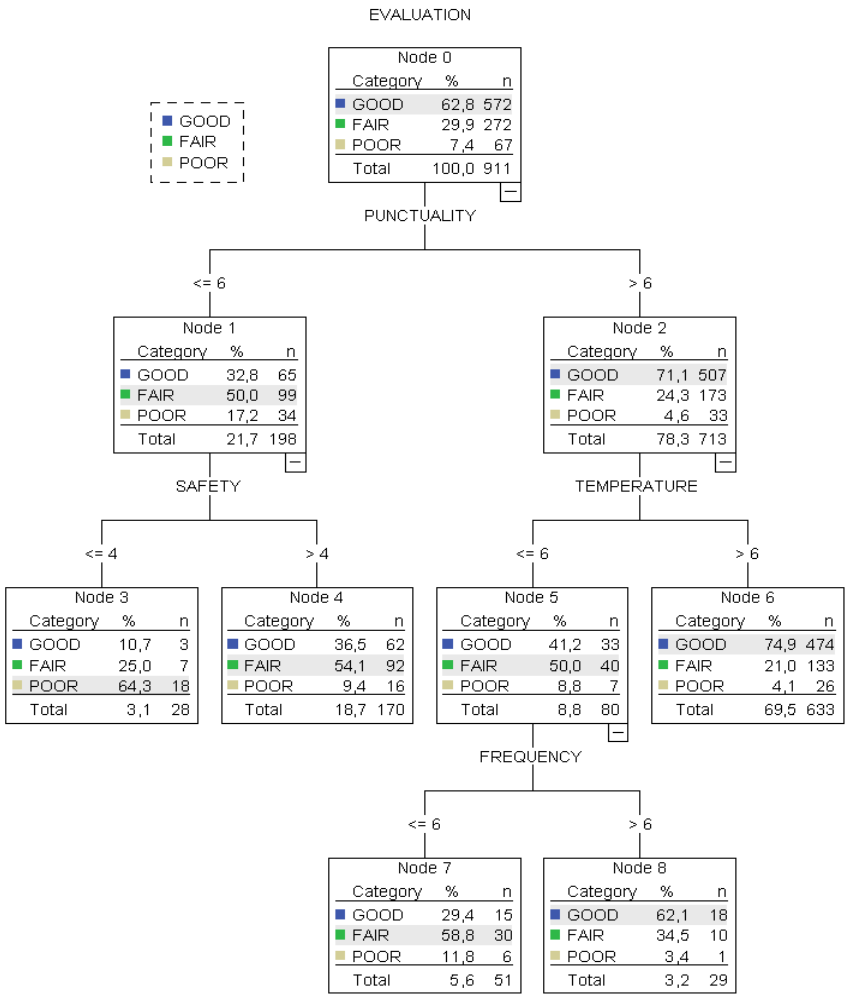


Figure 27. CART for users classified according to the travel reason (Studies)

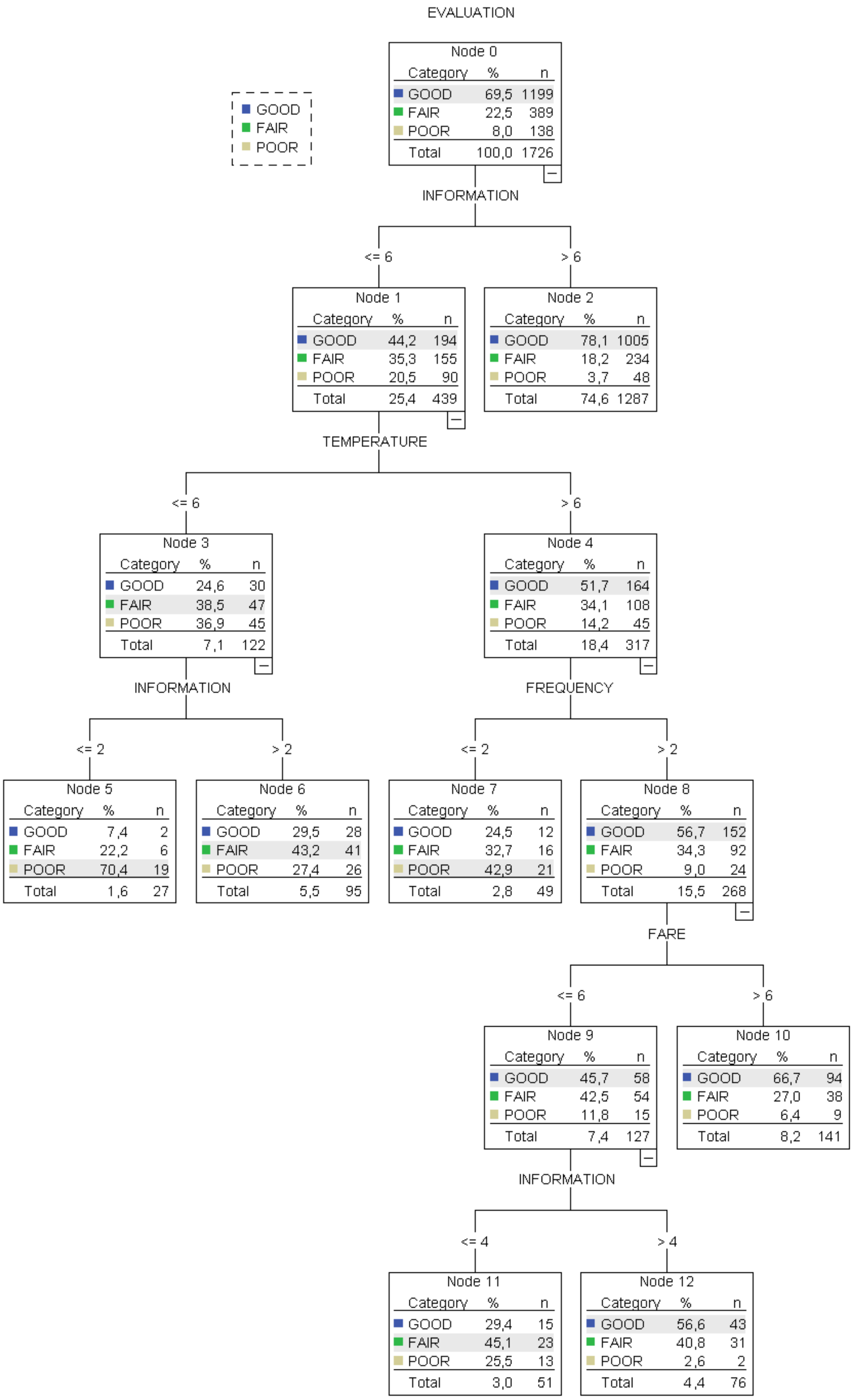


Figure 28. CART for users classified according to the travel reason (Others reasons)

5.2.2.2.5. Type of ticket market segments.

Figures 29 to 32 show the analysis based on the type of ticket. The individuals who used a Senior Citizen Pass discriminate their perception of quality in terms of Information, Accesibility and Courtesy. In general, most of those who use a Senior Citizen Pass have a good perception of Information (>4) and overall SQ (Node 2, in 91.2% of the cases). In the event that Information is rated negatively (≤4), the attribute Accesibility stands out in the model, predicting in Node 3 that when Information and Accesibility get low scores (≤4), the overall SQ is likely to be perceived as POOR (71.4%). This may be owing to the growing difficulties of mobility generally faced by the elderly. The discriminant factors for passengers who use Standard tickets, however, are Speed and Courtesy. Every time they use the service, these passengers need to communicate with the driver of the vehicle to buy a ticket (as opposed to the passengers who use different kinds of cards: Senior Citizen Pass or Consortium Pass) and so they tend to take into consideration the driver's friendliness more than those who do not exchange any word with the driver. In the case of those who use the Consortium Pass, the tree is more complex, owing to a range of attributes, including Information, Frequency, Proximity, Age and Temperature. This tree is very similar to the one built for the overall market, because most passengers (66.7%) use this kind of ticket. The accuracy attained in these models is 79.12% for Senior Citizen Pass, 63.65% for Standard ticket, 67.81% for Consortium pass, and 65.00% for Other tickets.

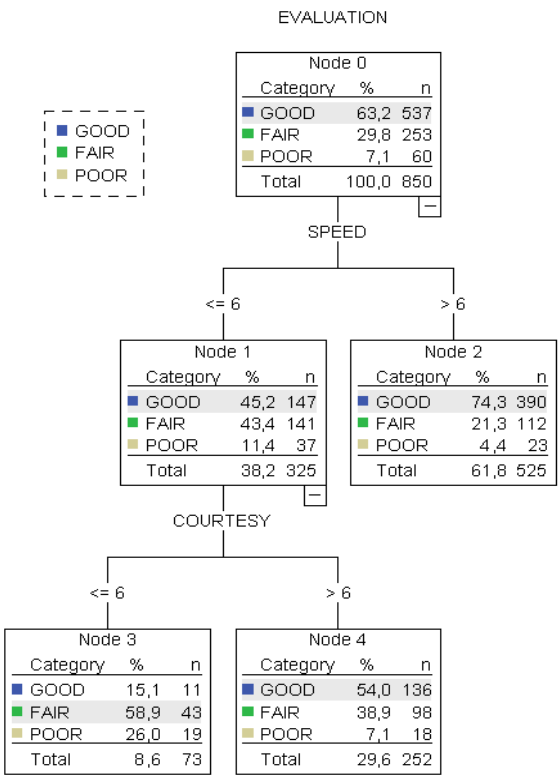


Figure 29. CART for users classified according to the type of ticket (Standard)

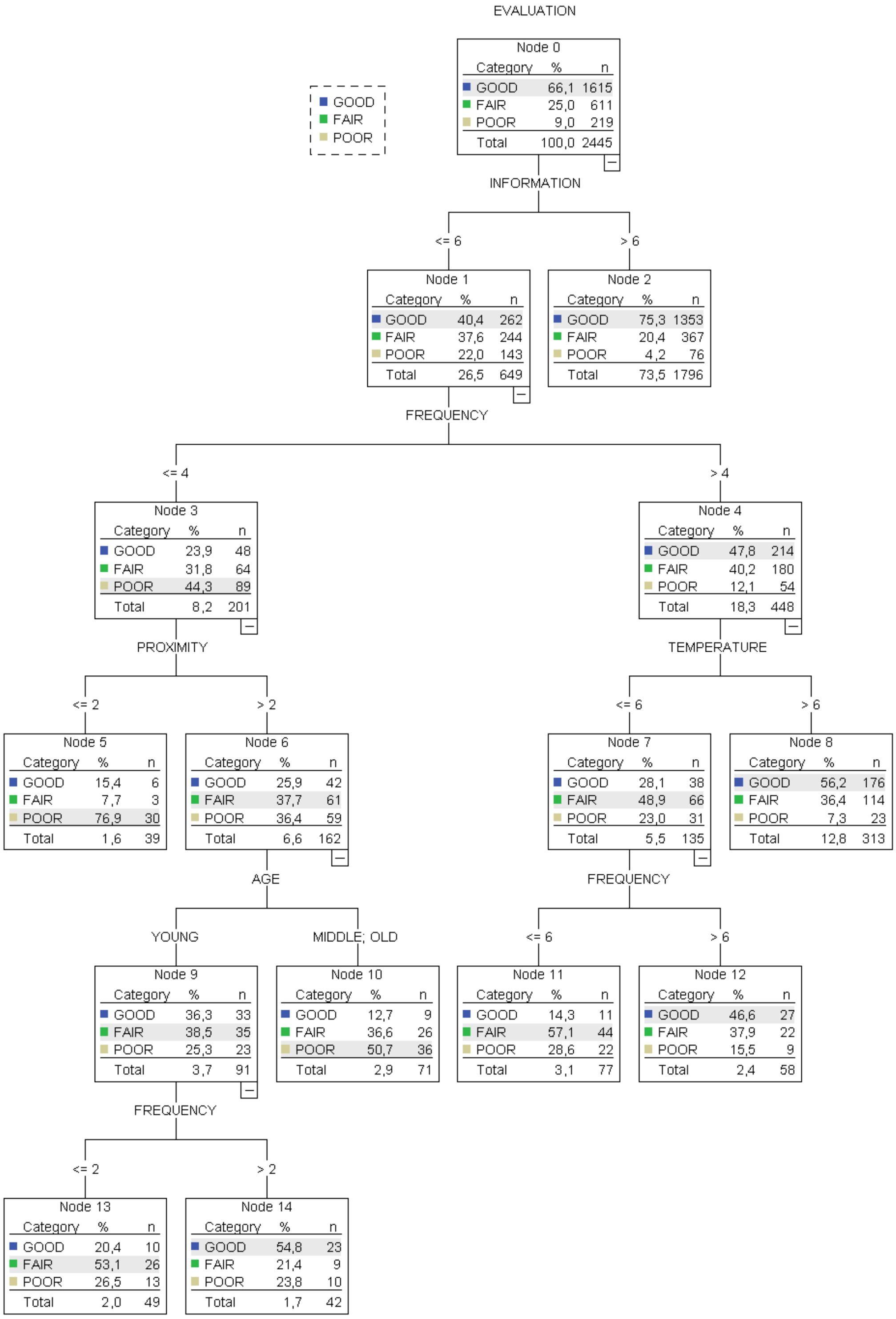


Figure 30. CART for users classified according to the type of ticket (Consortium pass)

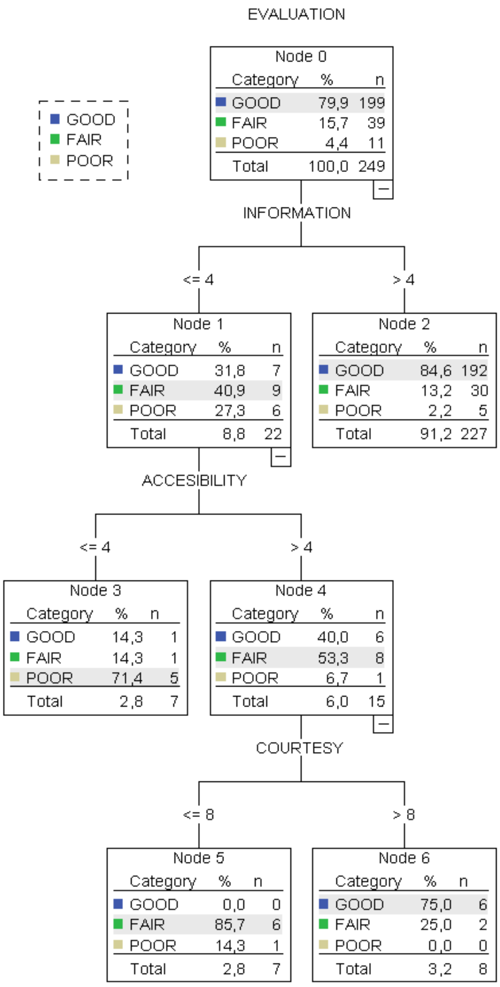


Figure 31. CART for users classified according to the type of ticket (Senior citizen pass)

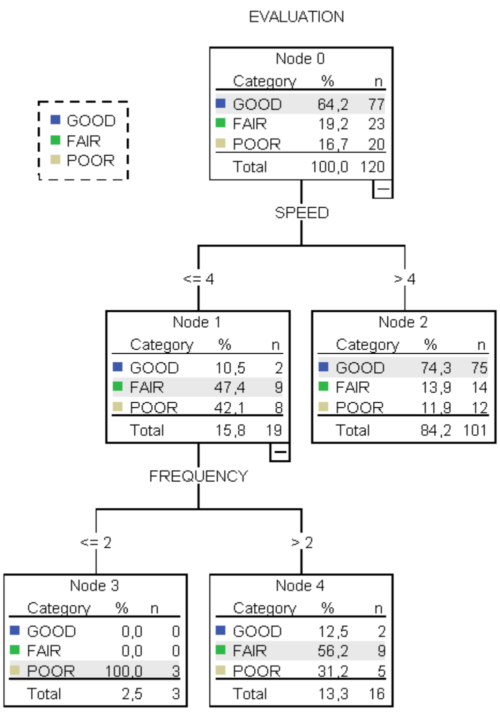


Figure 32. CART for users classified according to the type of ticket (Other tickets)

5.2.3. Importance of the variables

The importance of each independent variable on the model is extracted by the importance index (Kashani and Mohaymany, 2011). The results are compared with the importance frequencies stated by the passengers.

Table 17 shows the importance frequencies analyzed for different market segment (gender, age, travel reason, frequency of use and type of ticket) based on stated preferences. The five most important attributes for the overall market are still identified as the most important in almost all market segments. The same does not occur in three market segments: Elderly, Senior citizen pass and Other type of ticket. For elderly passengers and those who use a senior citizen pass, the three most relevant attributes for the overall market (Frequency, Fare and Punctuality) remain among the five most important attributes, but Safety and Accessibility also become highly important. In those market segments Safety is considered primordial, and Accessibility is among the top five attributes. With regards to "Other type of ticket", the attribute Information comes third in importance.

Table 17. Importance frequencies by market segment (stated importance based on CSS)

Market segment	Category	Variable	Frequency
OVERALL MARKET (max. obs. available 3,271)		FREQUENCY	1,700 (52.0%)
		FARE	1,605 (49.1%)
		PUNCTUALITY	1,577 (48.2%)
		SAFETY	986 (30.1%)
		SPEED	793 (24.2%)
GENDER (max. obs. available 3,271)	FEMALE (max. obs. available 2,229)	FREQUENCY	1,201 (53.9%)
		PUNCTUALITY	1,084 (48.6%)
		FARE	1,074 (48.2%)
		SAFETY	685 (30.7%)
		SPEED	521 (23.4%)
	MALE (max. obs. available 1,042)	FARE	531 (51.0%)
		FREQUENCY	499 (47.9%)
		PUNCTUALITY	493 (47.3%)
		SAFETY	301 (28.9%)
		SPEED	272 (26.1%)
AGE (max. obs. available 3,271)	YOUNG (max. obs. available 1,599)	FREQUENCY	860 (53.8%)
		FARE	852 (53.3%)
		PUNCTUALITY	844 (52.8%)
		SPEED	457 (28.6%)
		SAFETY	382 (23.9%)
	MIDDLE (max. obs. available 1,326)	FREQUENCY	702 (52.9%)
		FARE	654 (49.3%)
		PUNCTUALITY	598 (45.1%)
		SAFETY	428 (32.3%)
		SPEED	264 (19.9%)
	OLD	SAFETY	176 (50.9%)

	(max. obs. available 346)	FREQUENCY PUNCTUALITY FARE ACCESIBILITY	138 (39.9%) 135 (39.0%) 99 (28.6%) 82 (23.7%)
FREQUENCY OF USE (max. obs. available 3,271)	FREQUENT (max. obs. available 2,541)	FREQUENCY FARE PUNCTUALITY SAFETY SPEED	1,356 (53.4%) 1,303 (51.3%) 1,262 (49.7%) 726 (28.6%) 620 (24.4%)
	SPORADIC (max. obs. available 730)	FREQUENCY PUNCTUALITY FARE SAFETY SPEED	344 (47.1%) 315 (43.2%) 302 (41.4%) 260 (35.6%) 173 (23.7%)
TRAVEL REASON (max. obs. available 3,271)	OCCUPATION (max. obs. available 920)	FREQUENCY PUNCTUALITY FARE SAFETY SPEED	496 (53.9%) 472 (51.3%) 442 (48.0%) 250 (27.2%) 198 (21.5%)
	STUDIES (max. obs. available 796)	FARE PUNCTUALITY FREQUENCY SPEED SAFETY	457 (57.4%) 436 (54.8%) 432 (54.3%) 241 (30.3%) 175 (22.0%)
	OTHERS (max. obs. available 1,555)	FREQUENCY FARE PUNCTUALITY SAFETY SPEED	772 (49.6%) 706 (45.4%) 669 (43.0%) 561 (36.1%) 354 (22.8%)
TYPE OF TICKET (max. obs. available 3,271)	STANDARD (max. obs. available 781)	PUNCTUALITY FREQUENCY FARE SAFETY SPEED	408 (52.2%) 396 (50.7%) 337 (43.1%) 274 (35.1%) 207 (26.5%)
	CONSORTIUM PASS (max. obs. available 2,169)	FREQUENCY FARE PUNCTUALITY SAFETY SPEED	1,166 (53.8%) 1,160 (53.5%) 1,058 (48.8%) 569 (26.2%) 524 (24.2%)
	SENIOR CITIZEN PASS (max. obs. available 232)	SAFETY FREQUENCY PUNCTUALITY ACCESIBILITY FARE	127 (54.7%) 91 (39.2%) 82 (35.3%) 65 (28.0%) 62 (26.7%)
	OTHER (max. obs. available 89)	FREQUENCY FARE INFORMATION PUNCTUALITY SAFETY	45 (50.6%) 39 (43.8%) 25 (28.1%) 23 (25.8%) 22 24.7%)

Therefore, regarding the importance stated by the passengers, no significant differences are observed among the different market segments, contrary to the results from earlier studies (Andreassen, 1995; dell'Olio et al., 2010; Ganesan-Lim et al., 2008). Once again, these results point to the limitations of using stated importance to identify the importance of each attribute (Weinstein, 2000).

Table 18 shows the normalized importance of the variables deduced from each one of the models developed. For simplicity's sake, the table only shows the five most important variables in each case.

Table 18. Derived importance by market segment based on service quality perception (bus transit service 2008-2011)

Market segment	Category	Variable	Normalized
OVERALL MARKET (n. obs. 3,664; accur. rate 68.56%)		PUNCTUALITY	100.0%
		TEMPERATURE	92.3%
		INFORMATION	91.3%
		FREQUENCY	86.0%
		SAFETY	70.3%
GENDER (n. obs. 3,664)	FEMALE (n. obs. 2493; accur. rate 65.26%)	PUNCTUALITY	100.0%
		INFORMATION	70.6%
		FREQUENCY	64.7%
		TEMPERATURE	63.2%
		CLEANLINESS	57.7%
	MALE (n. obs. 1171; accur. rate 71.98%)	INFORMATION	100.0%
		PUNCTUALITY	90.4%
		SAFETY	81.1%
		FREQUENCY	78.6%
		COURTESY	72.2%
AGE (n. obs. 3,664)	YOUNG (n. obs. 1,809; accur. rate 64.73%)	PUNCTUALITY	100.0%
		SAFETY	77.8%
		SPEED	55.6%
		PROXIMITY	37.9%
		COURTESY	34.5%
	MIDDLE (n. obs. 1479; accur. rate 68.15%)	FREQUENCY	100.0%
		INFORMATION	99.9%
		PUNCTUALITY	87.8%
		SPACE	85.7%
		TEMPERATURE	79.8%
	OLD (n. obs. 376; accur. rate 78.98%)	INFORMATION	100,0%
		COURTESY	45,8%
		PROXIMITY	37,5%
		ACCESIBILITY	37,5%
		SPEED	32,5%
FREQUENCY OF USE (max. obs. available	FREQUENT (n. obs. 2870; accur.	PUNCTUALITY	100.0%
		INFORMATION	94.2%
		FREQUENCY	92.1%
		SPEED	74.0%

3,271)	rate 68.29%)	SAFETY	71.4%
	SPORADIC (n. obs. 794; accur. rate 69.77%)	PUNCTUALITY	100.0%
		SAFETY	64.5%
		CLEANLINESS	27.1%
		COURTESY	26.3%
		TEMPERATURE	20.0%
TRAVEL REASON (n. obs. 3,664)	OCCUPATION (n. obs. 1027; accur. rate 65.86%)	FREQUENCY	100.0%
		INFORMATION	93.4%
		PUNCTUALITY	89.9%
		SPACE	80.8%
		CLEANLINESS	80.7%
	STUDIES (n. obs. 911; accur. rate 67.67%)	PUNCTUALITY	100.0%
		TEMPERATURE	90.1%
		SPACE	89.7%
		SAFETY	71.1%
		INFORMATION	51.3%
	OTHERS (n. obs. 1726; accur. rate 69.58%)	INFORMATION	100.0%
		PUNCTUALITY	71.3%
		CLEANLINESS	65.2%
		SAFETY	61.2%
		COURTESY	51.6%
TYPE OF TICKET (max. obs. available 3,271)	STANDARD (n. obs. 850; accur. rate 63.65%)	COURTESY	100.0%
		SPEED	83.0%
		PUNCTUALITY	76.0%
		FREQUENCY	72.7%
		INFORMATION	71.8%
	CONSORTIUM PASS (n. obs. 2445; accur. rate 67.81%)	PUNCTUALITY	100.0%
		INFORMATION	85.9%
		SAFETY	82.6%
		TEMPERATURE	64.3%
		SPACE	62.8%
	SENIOR CITIZEN PASS (n. obs. 249; accur. rate 79.12%)	INFORMATION	100.0%
		ACCESIBILITY	82.8%
		SPACE	61.4%
		CLEANLINESS	58.1%
		COURTESY	55.5%
	OTHER (n. obs. 120; accur. rate 65.00%)	SPEED	100.0%
		ACCESIBILITY	43.5%
		PUNCTUALITY	23.8%
		SPACE	22.8%
		FREQUENCY	20.8%

Punctuality is the most important attribute on SQ in metropolitan bus transport for the overall market (Table 18). A number of authors (dell'Olio et al., 2010; Eboli and Mazzulla, 2010), who have analyzed SQ for bus transport, have also identified Punctuality as one of the attributes with the greatest impact on overall SQ. The variables Temperature, Information and Frequency are also identified as having a lot of weight on overall market. Two of these (Temperature and Frequency) coincide with the variables obtained by dell'Olio et al. (2010) in their work. In that case, they did not evaluate Temperature as a separate attribute of the service. Instead, they used overall comfort of the service as a variable that can encompass Temperature as well. It would be reasonable to suppose that travel comfort is a relevant attribute on long journeys, as in the case of metropolitan transport.

Moreover, Punctuality was found among the three most important variables in all market segments, with the exception of elderly people and those who used a Senior Citizen Pass (both groups are highly correlated). Most of these people are retired and do not have to comply with a schedule for work or studies, so they attach more importance to Information, Accesibility or Courtesy.

Information is another attribute that is highly important in most categories, although it is less important to young people and sporadic passengers. Young people may not attach much importance to Information because they are skilled at using new technologies and/or the information panels for travelers. Sporadic passengers only use PT occasionally; what matters the most important to them is Punctuality, with other attributes being considerably less important.

There are three attributes that are repeated many times (7) among the five most important attributes for various market segments: Frequency, Safety and Courtesy. Frequency is identified as one of the five most important attributes for men and women, middle-aged passengers, frequent users, users who travel for working reasons and passengers who use standard or other tickets. Safety is identified as one of the five most important attributes for males, young people, frequent and sporadic users, passengers who travel for studies or other reasons, and for consortium pass users. The profile of users who attach considerable value to Courtesy would be male, young or old, use the service sporadically with others reasons for travel, and who uses a standard ticket or senior citizen pass. It seems logical that elderly or infrequent users, those that usually buy their ticket on board and therefore they must interact with the driver, to rate Courtesy as one of the most important aspects.

There is another group of attributes related to travel comfort (Space, Temperature and Cleanliness) that is repeated in five or six market segments. This too, is natural, considering that these are metropolitan trips that tend to have longer itineraries. Another aspect worth mentioning is the fact that some tributes are considered highly important only for some specific market segments; for example, Accessibility attains importance values of 82.8% for Senior Citizen Pass users, but it is not among the five top attributes in the remaining market segments.

Finally, Table 18 shows that Fare is not among the most important five attributes for the overall market nor any of the market segments. This contrasts with the results shown in Table 17, where Fare is among the three most frequent attributes, and also it was one of the attributes with the lowest average SQ perception. To the contrary, it also shows that Information is identified as one of the five most important attributes for the overall market and for eleven market segments, and yet it is among the five most frequent attributes in only one market segment (Table 17). Information is also one of the three attributes that has the lowest average SQ perception. Such differences may be due to the following reasons:

- When people are asked to rate Fare, or the importance of Fare, they frequently rate it as high and as very important for SQ. Despite this, when the weight of variable Fare is obtained using models based on the overall SQ, fares are found to be less important that other variables that have a much greater impact (frequency, punctuality, safety, etc.) (dell'Olio et al., 2010; de Oña et al., 2012).

- The contrary appears to occur with Information. When users are asked about this variable directly, they do not usually rate it very highly (de Oña et al., 2012), and yet, when it is inferred from the models based on overall SQ, the importance of Information increases (Andreassen, 1995; de Oña et al., 2012; Eboli and Mazzulla, 2010).

Table 18 shows which attributes in the categories of users studied will lead to service quality, thus providing valuable information for transport operators and managers.

5.3. Study Case 3: Decision Trees for the rail public service

The rail public service experimental context was used to investigate about the heterogeneity among users, by proposing an analysis of service quality conducted on the basis of users' perceptions, expressed in terms of satisfaction assigned to various service characteristics. The CART model was applied to investigate about the perceptions of different groups of users on the services provided by a rail operator of the North of Italy, with the final aim to identify which are the characteristics mostly influencing the overall service quality perceived by the different kinds of users. Therefore, the main differences in users' perceptions about the services can be verified.

The different market segments were distinguished according to four criteria: the type of user in terms of the purpose of the journey by considering commuters travelling for working, commuters travelling for studying, and people travelling for other purposes; the type of the day of the journey, by distinguishing people travelling during the week, in days before a holiday, and in holidays; the frequency of the use of the service, which allows us to make the difference between users travelling daily, users travelling weekly, and people travelling occasionally; the time of the day, according to which users were classified in users travelling in the off-peak hours, in the morning peak hours, afternoon peak hours, and evening peak hours.

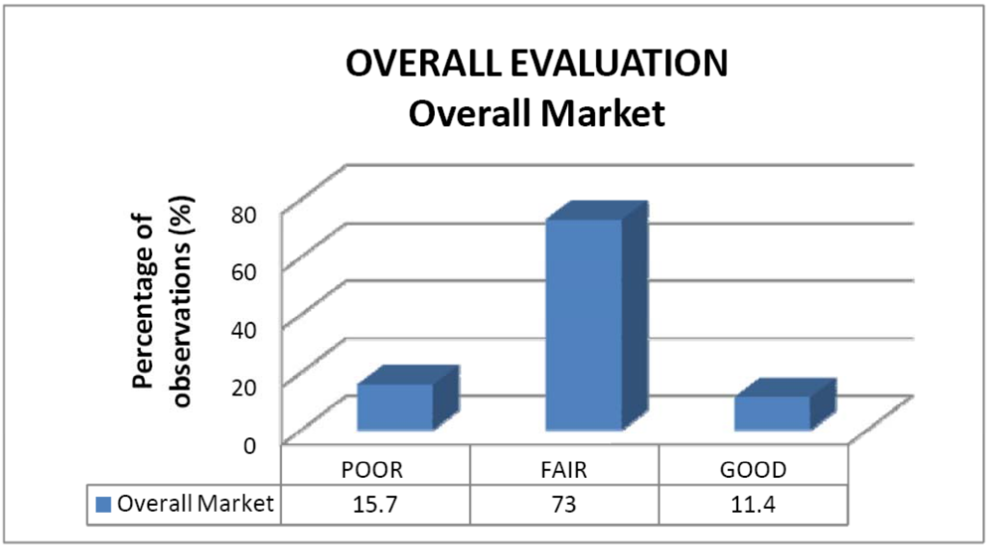
Data from a CSS conducted in May of 2012 in 9 suburban lines connecting towns of the hinterland of Milan were used. Thirteen different CART decision trees were built – one for each market segment – and another CART was built for the overall market. The 27 SQ attributes were the variables used as predictors. The findings arisen from the application of the CART methodology could be very useful for the operators of the service and the policy makers to identify the strategy to be adopted for the improvement of the service by considering the different market segments fo users.

5.3.1. Data preparation

In this experimental context, the scale of the target variable, as well as the scale of the predictors of the models (in this case only were used the SQ attributes), were recoded in a reduced 3-point semantic scale in order to find more applicable decision rules. The semantic scale comprised the rates from 1 to 4 as POOR, from 5 to 7 as FAIR, and from 8 to 10 as GOOD.

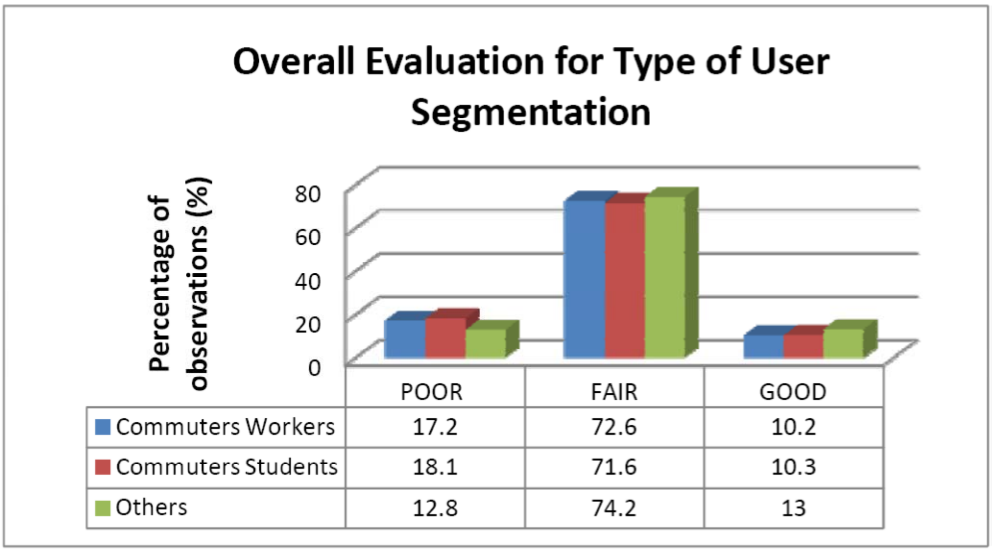
Figures 33 to 37 display the distribution of the Overall SQ perceptions recoded into the three point semantic scale. Imbalance data are shown, with several cases for the class of the target variable defined as FAIR, and few cases for the other two classes (POOR and GOOD).

Figure 33 presents the distribution of observations about the Overall Evaluation for the Overall Market. High percentage is reached at the FAIR class, and low values for the POOR and GOOD cases, as was mentioned before. This distribution is followed across the different market segments under study. The market segments that achieved the lowest number of observations classified as POOR were the passegers travelling in Holidays and those that used the service Occasionally, with percentages of 10.8% and 11.6% respectively. On the other hand, the segments that achieved the highest proportion of observations considered as POOR were the Commuters Students (18.1%), the passengers travelling the days before holidays (18.8%), and those that travel with a daily frequency (17.6%). Regarding the passengers that perceive the Overall Evaluation as GOOD, the passengers that travel in Holidays were the ones with the highest percentage, with value of 16,7%. The market segments that less perceived the Overall Evalaution as GOOD were the Communters Workers (10.2%), Commuters Students (10.3%), passengers travelling the days before holidays (8%) and passengers travelling in the Evening Peak Hour. Concerning the proportion of FAIR observations, its value was high across the different market segments and also in the overall market. It ranged from a 71.6% for Commuter students and passengers travelling in the morning peak hour, to a 76.3% for passengers travelling in the evening peak hour.



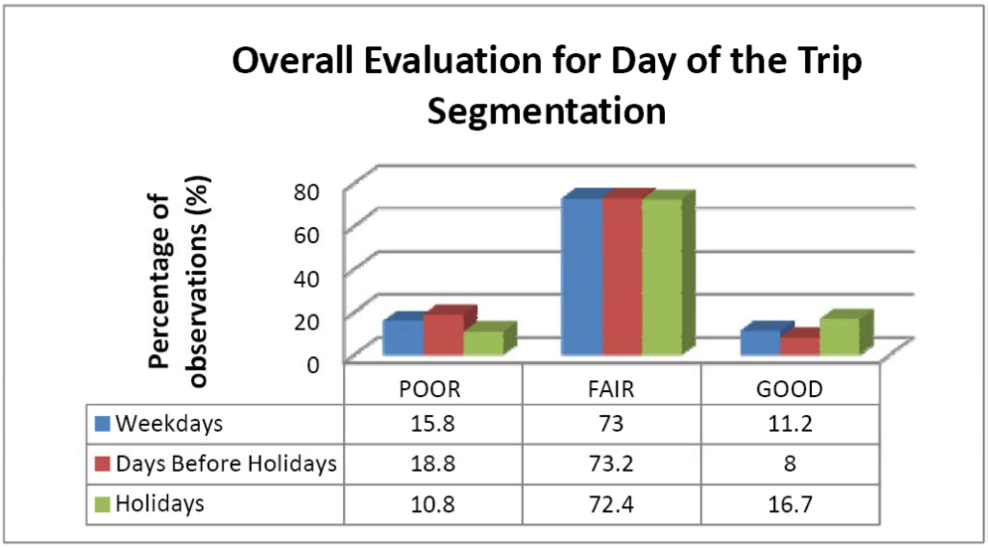
	POOR	FAIR	GOOD
Overall Market	15.7	73	11.4

Figure 33. Overall Evaluation for the Overall Market



	POOR	FAIR	GOOD
Commuters Workers	17.2	72.6	10.2
Commuters Students	18.1	71.6	10.3
Others	12.8	74.2	13

Figure 34. Overall Evaluation according to Type of User Segmentation



	POOR	FAIR	GOOD
Weekdays	15.8	73	11.2
Days Before Holidays	18.8	73.2	8
Holidays	10.8	72.4	16.7

Figure 35. Overall Evaluation according to Day of the Trip Segmentation

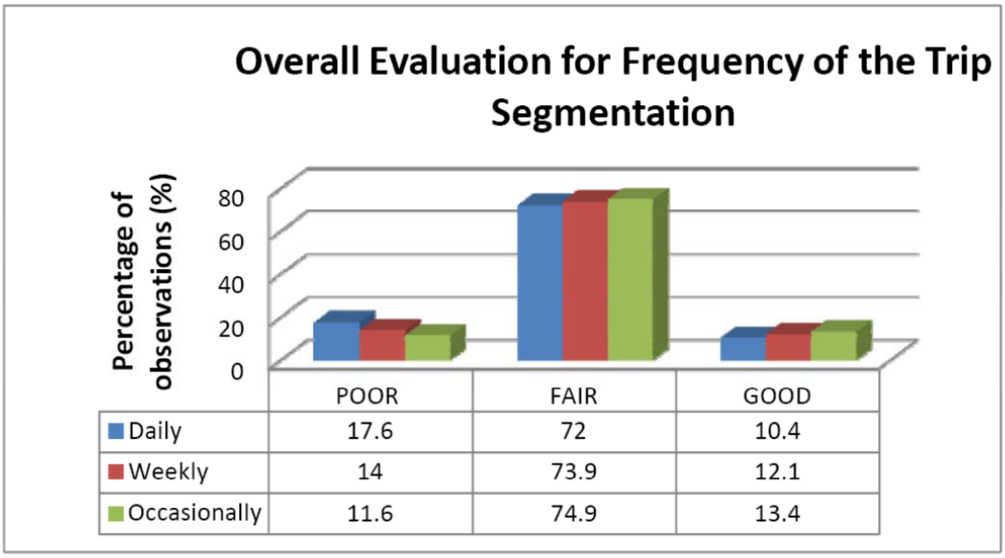


Figure 36. Overall Evaluation according to Frequency of the Trip Segmentation

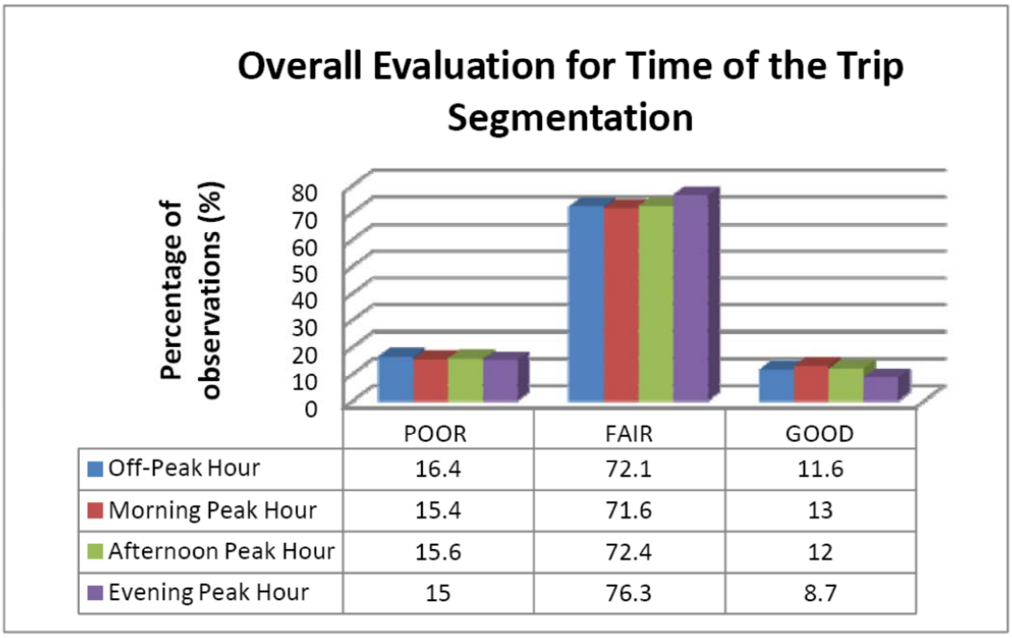


Figure 37. Overall Evaluation according to Time of the Trip Segmentation

5.3.2. Decision Trees

Different groups of users were considered in order to identifying differences in the perceptions of service quality, and the characteristics describing the service. So, in the following the results of different applications will be described. Firstly, CART is applied to all the 7,333 passengers of the analyzed suburban lines. Afterwards, CART is applied to the users classified according to the four different criteria.

For all the models, the 27 attributes describing the service (Item1-Item27) were used as independent variables. For all the groups, CART used a 10-fold cross-validation of the sample, which gave high accuracy ratio of the categorization of the variable class from about 76.00% to 79.00%.

5.3.2.1. CART for the overall market

The accuracy rate for the overall market of the suburban service is 78.15%. The tree produced 5 levels, 21 nodes and 11 terminal nodes. The variable that splits the root node and obtains the maximum purity of the two child nodes is the Item12 "Fare/service ratio" (Figure 38).

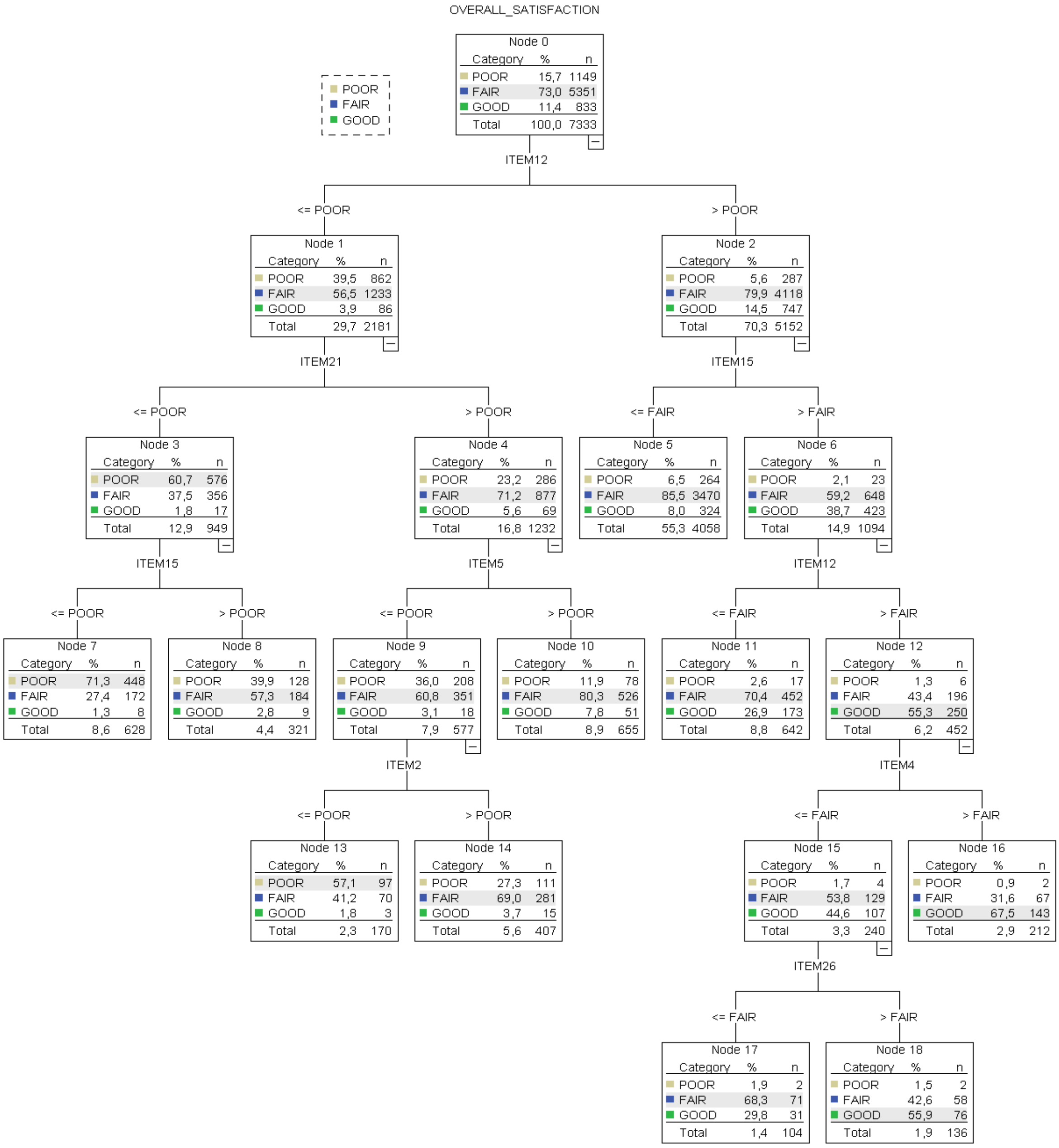


Figure 38. CART for suburban lines

On the left branch of the tree there are the passengers having a Poor satisfaction with this item; it is predicted in the 6 terminal nodes of the branch (8, 10, 13, 14, 15, and 16), where the "Overall satisfaction" will be POOR or FAIR. This implies that when a passenger has a Poor satisfaction with "Fare/service ratio", his/her overall satisfaction with the service will never be GOOD. Under Node1, the tree grows according to different splitters as Item21 “Information at stations”, Item15 “Regularity of runs”, Item5 “Cleanliness of seats”, and Item2 “Personal security on board”.

On the other hand, on the right branch, there are the passengers having a Fair or Good satisfaction with “Fare/service ratio”. In this case, there are 5 terminal nodes (5, 11, 16, 17 and 18) and their prediction of the “Overall satisfaction” is FAIR or higher. If Item 15 “Regularity of runs ” and Item 12 “Fare/service ratio” are considered with Good satisfaction, the probability to have a GOOD “Overall satisfaction” increases, and two terminal nodes classified as GOOD can be achieved (Node 16 and Node 18, with probabilities of 67.5% and 55.9% respectively).

5.3.2.2. CART for the different market segments

The CART methodology was also applied to different groups of users divided according to four criteria: the type of the user depending on the purpose of the journey, the type of the day of the journey, the frequency of the use of the service, the time of the day of the journey. The size of each group is reported in Table 19. Most of the sample travels for purposes different from working or studies (40.4%), in a weekday (84.6%), daily (61.9%), in the off-peak hours (30.7%).

Table 19. Size of each group of users (Rail public service)

Criteria of Classification	*Category of User*	*Number of Observations*	*% of the Sample*
6. Type of User	Commuter Workers	2,371	32.3%
	Commuter Students	2,000	27.3%
	Others	2,962	40.4%
7. Day of the Trip	Weekdays	6,207	84.6%
	Days before Holidays	600	8.2%
	Holidays	526	7.2%
8. Frequency of the Trip	Daily	4,542	61.9%
	Weekly	1,064	14.5%
	Occasionally	1,727	23.6%
9. Time of the Trip	Off-Peak Hour	2,250	30.7%
	Morning Peak Hour	1,382	18.8%
	Afternoon Peak Hour	2,154	29.4%
	Evening Peak Hour	1,547	21.1%

5.3.2.2.1. Type of user market segments.

Three different trees were built according to the type of user: “Commuter Workers”, “Commuter Students” and “Others”. The precision achieved in each of these trees is high and similar to the one obtained with the global sample (78.87%, 77.60% and 77.28% respectively).

The variable splitting the root node is different in each one of the three cases (see Figures 39 to 41). For the tree built for the “Commuter Workers” the splitter is the Item 12 “Fare/service rate” as for the total sample; for the group of “Commuter

Students" the root node is divided into two child nodes based on the Item14 "Punctuality of runs", while the tree built for "Others" starts growing with the Item11 "Windows and doors working".

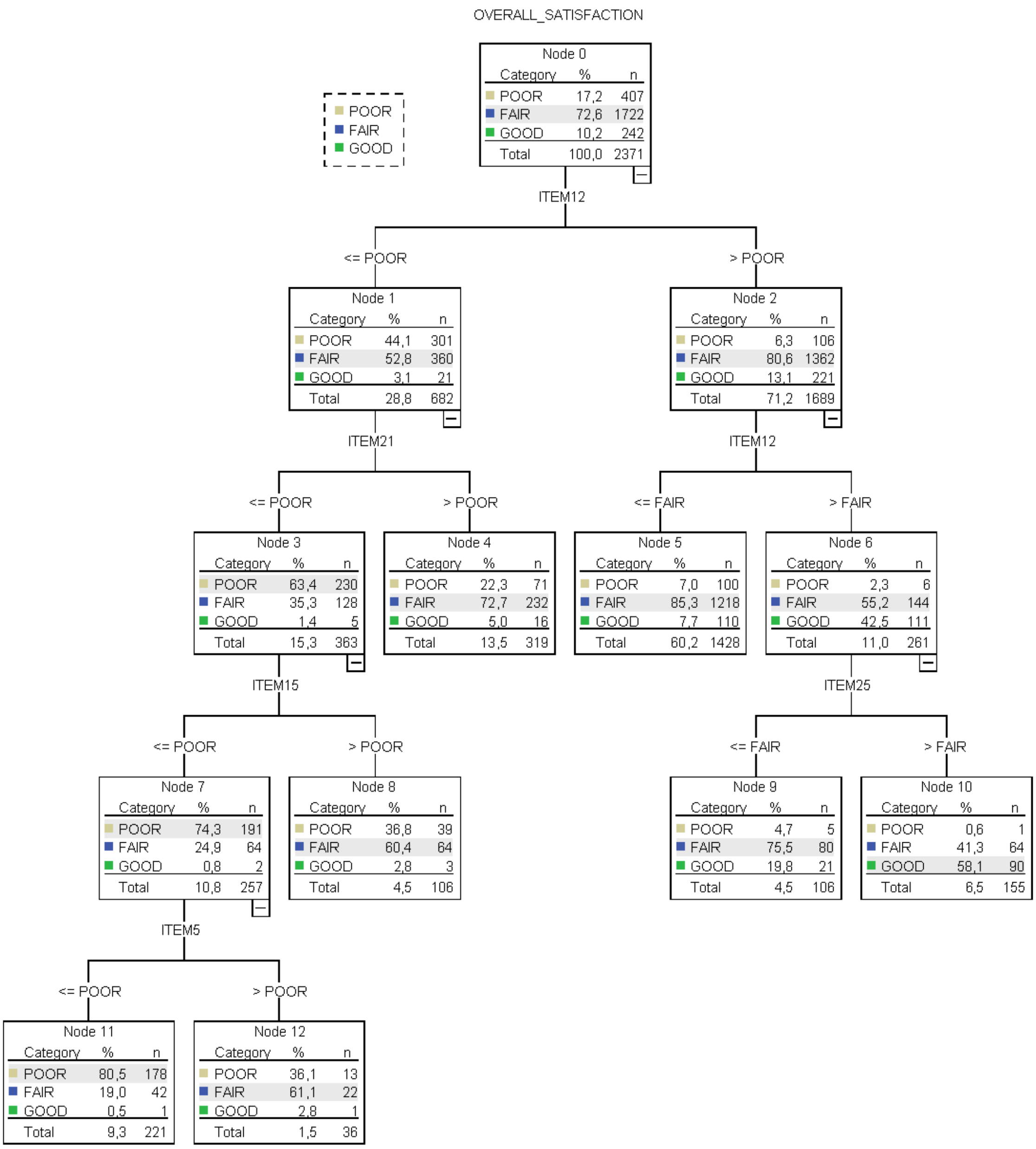


Figure 39. CART for users classified according to the type of user ("Commuter Workers")

In Figure 39, the tree derived for the "Commuter Workers" group is shown. When a group of variables (Item12 "Fare/service rate", Item21 "Information at stations", Item15 "Regularity of runs", and Item5 "Cleanliness of seats") are perceived with a Poor satisfaction, the probability of having a POOR "Overall satisfaction" is quite high (80.5% at Node11). On the other hand, when Item12 "Fare/service rate" and Item25 "Courtesy and competence on board" have a Good satisfaction, also the "Overall satisfaction"will be GOOD (Node10).

In the case of "Commuter Students" (Figure 40), for achieving a GOOD overall satisfaction with the service, not only the "Punctuality of the runs" must be Fair or Good, but also the Item12 "Fare/service rate" and the Item27 "Courtesy and competence on the stations" must be perceived as Good (Node12, 61.7%). And the tree built for "Others" (Figure 41) divides at the left branch the passengers that have a POOR or FAIR "Overall Satisfaction" with the service, and at the right branch, those with a FAIR or higher "Overall Satisfaction".

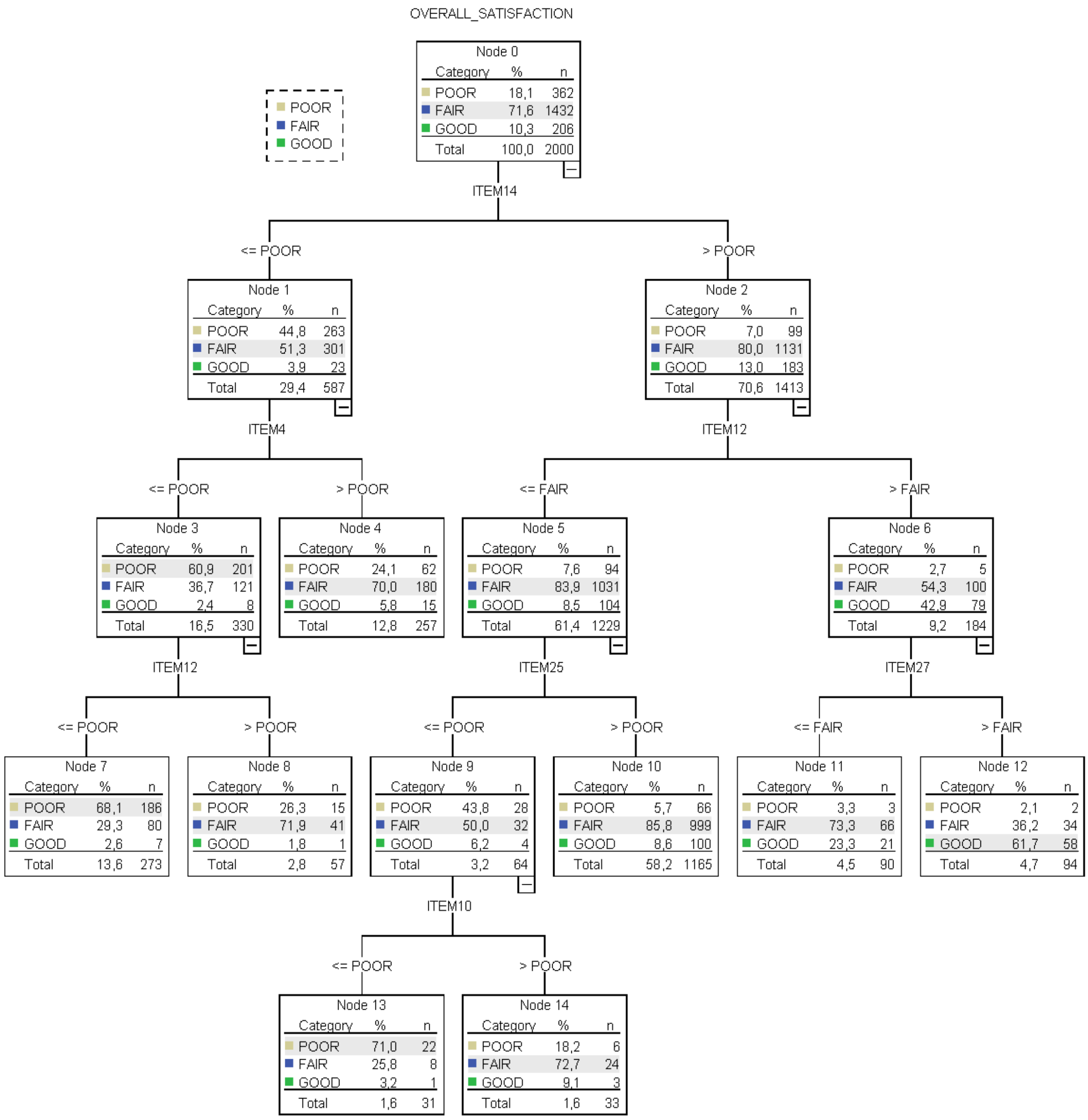


Figure 40. CART for users classified according to the type of user ("Commuter Students")

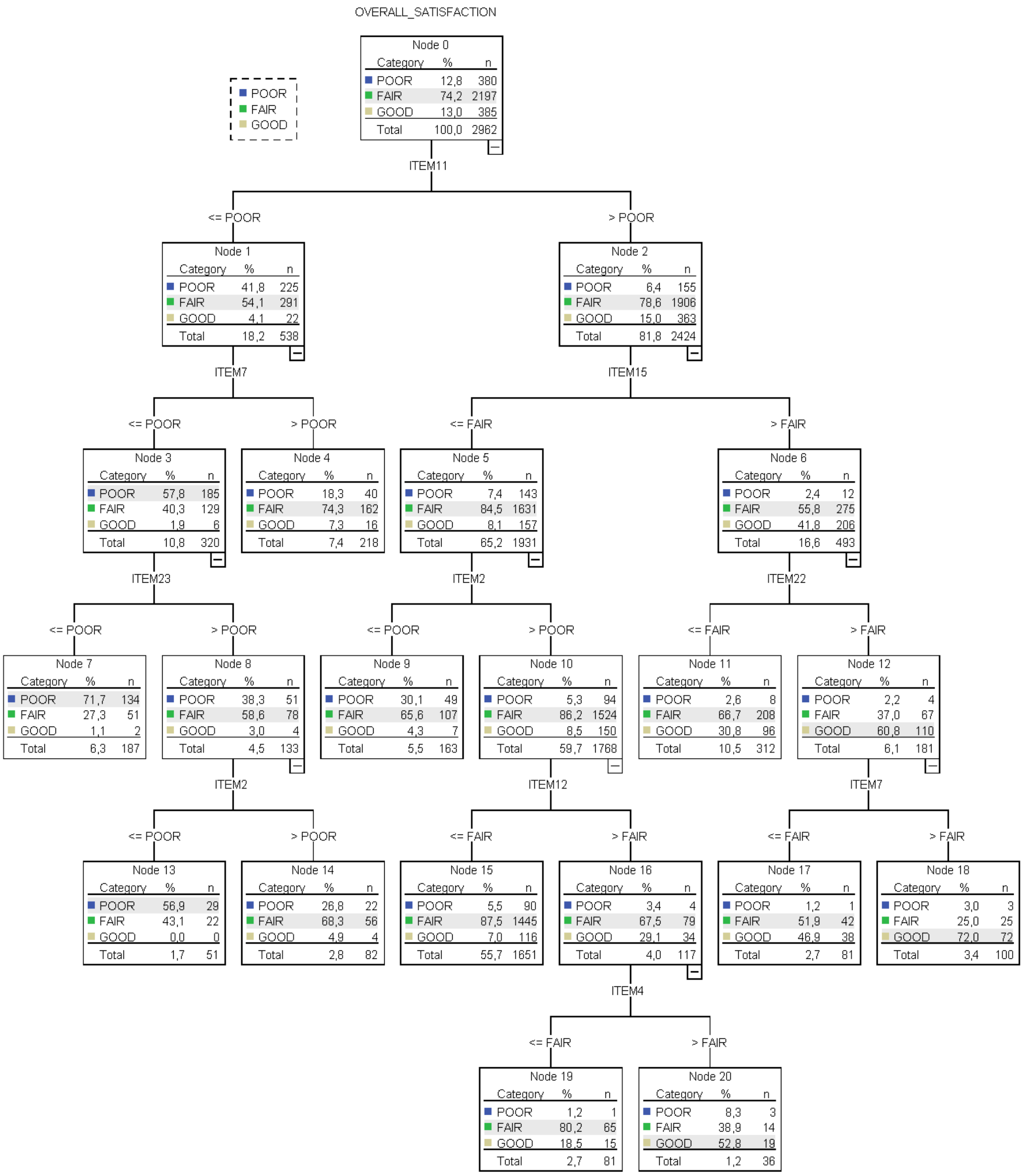


Figure 41. CART for users classified according to the type of user ("Others")

5.3.2.2.2. Type of day of the journey market segments.

The trees obtained by classifying the users according to the different type of day of the journey have high and similar precisions of about 77% (see Table 20). Item12 "Fare/service rate" is the splitter for the root node of "Working days".

We want to remark that the group "Working days" includes most part of the sample (84.6%), and the variable splitting the root node is also "Fare/service rate" in the tree built with the global sample (Figure 42). This splitter divides the tree into two branches. On the left, there are the passengers with a Poor satisfaction with the "Fare/service rate", 7 terminal nodes (7, 10, 14, 15, 16, 19 and 20) were produced predicting a POOR or FAIR "Overall satisfaction". On the right branch, when Item12 "Fare/service rate" have a Good satisfaction, then the "Overall satisfaction" will be FAIR or GOOD.

The tree built for the "Days before holiday" (see Figure 43) begins growing with Item27 "Courtesy and competence in stations". In this tree a little part of the sample has a GOOD overall satisfaction (only 8%) and no node predicts a GOOD satisfaction. For the group "Holidays" (Figure 44), the tree starts splitting with the Item12 "Fare/service rate". This attribute has been also identified as the most important attribute (Table 20).

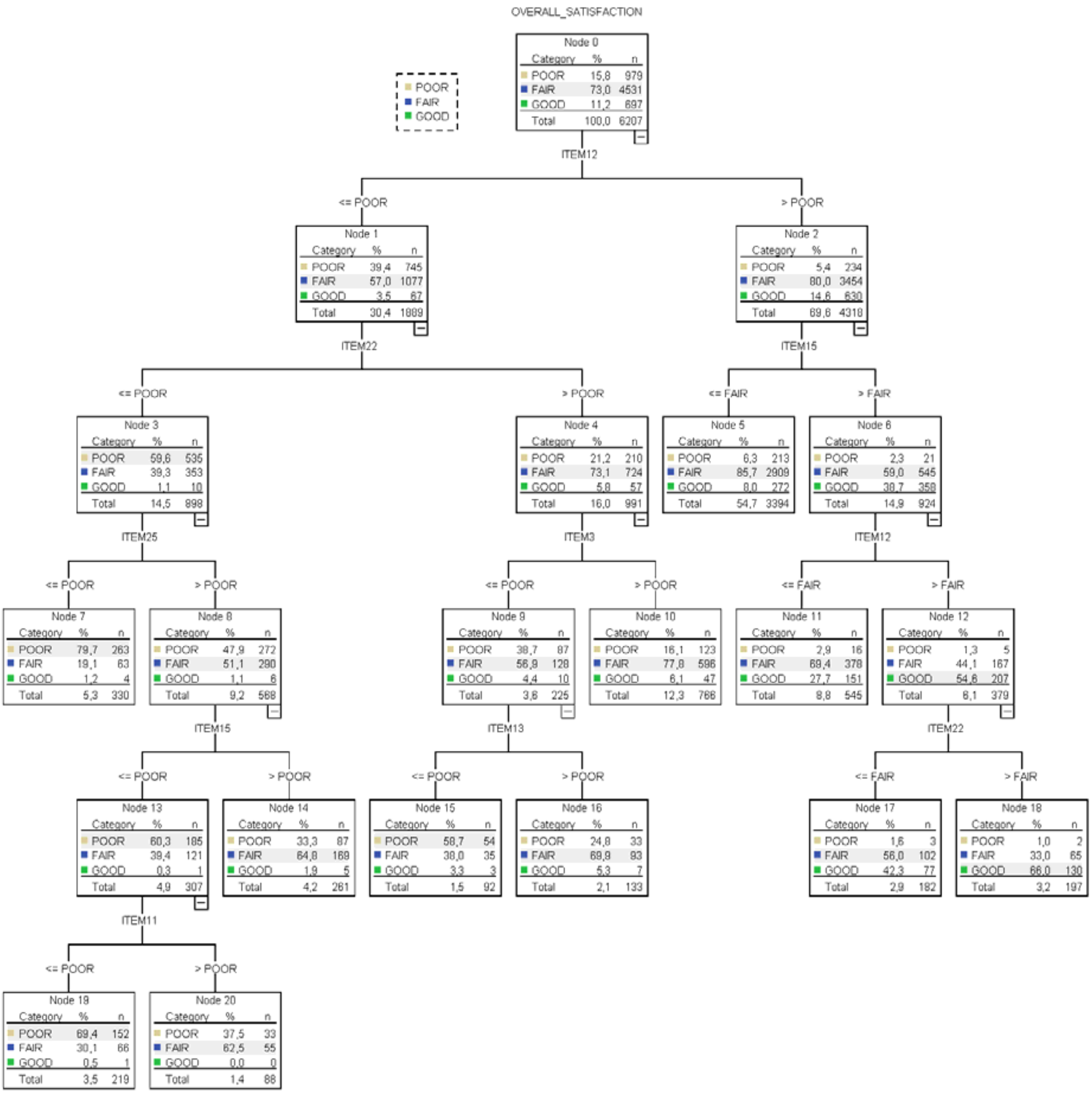


Figure 42. CART for users classified according to the day of the trip ("Working days")

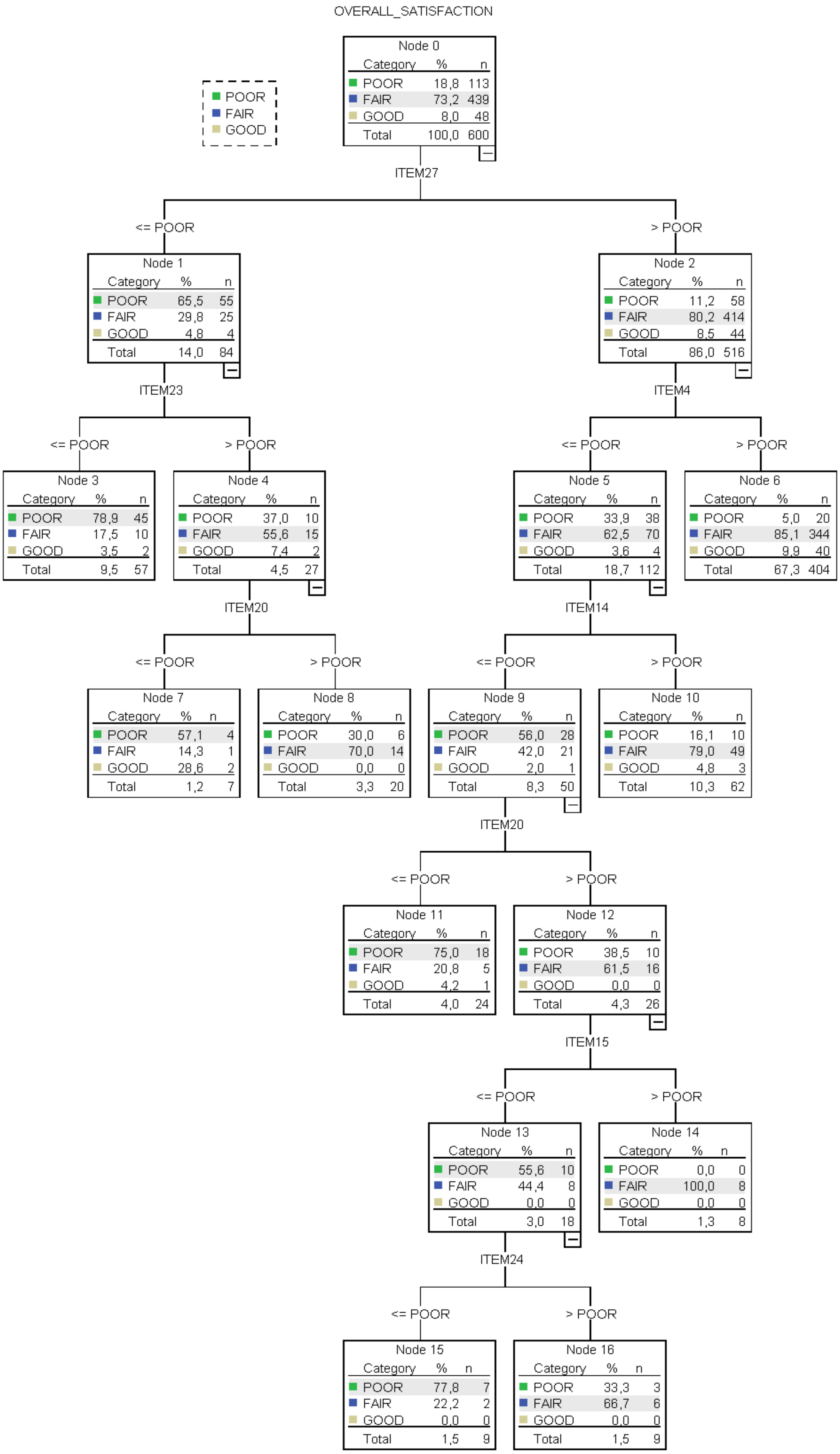


Figure 43. CART for users classified according to the day of the trip ("Days before holiday")

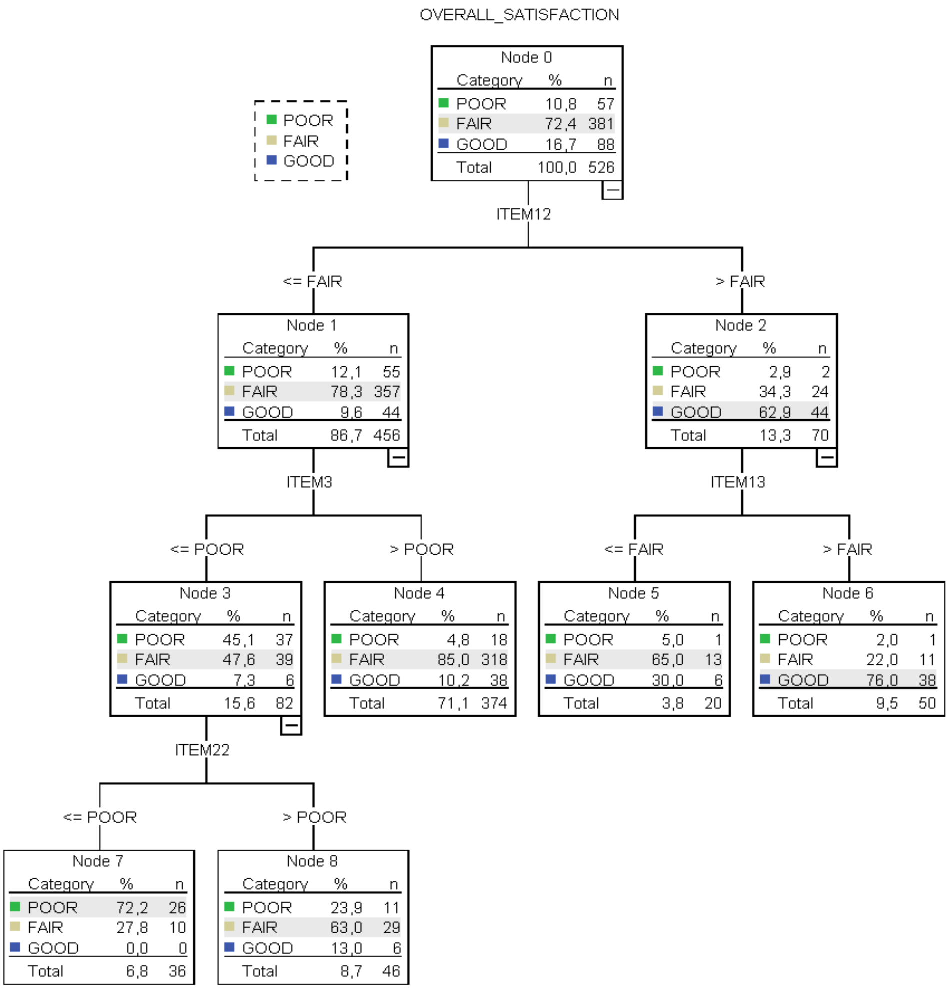


Figure 44. CART for users classified according to the day of the trip ("Holidays")

5.3.2.2.3. Frequency of the journey market segments.

All the trees obtained by classifying the users according to the different frequency of the journey have high precision rate, higher than 77%. The variable splitting the root node is different in each case. For users daily travelling, the splitter is Item12 "Fare/service rate", for users weekly travelling is Item15 "Regularity of runs", and for users who occasionally travel, the variable achieving the best classifications is the Item25 "Courtesy and competence on board".

For the "Daily"group, which is the largest group (61.9% of the users), the tree produced 5 levels, 25 nodes and 13 terminal nodes (Figure 45). When Item12 "Fare/service rate", and also Item22 "Information on board" have a Poor satisfaction, the probability of having a POOR "Overall satisfaction" increases. If also Item25 "Courtesy and competence on board" has a Poor satisfaction, terminal node 7 is created with a prediction of the "Overall satisfaction" as POOR. On the other hand, when Item12 "Fare/service rate" and Item25 "Courtesy and competence on board" are Good, the probability of having a GOOD overall satisfaction is higher. The accuracy indicator for this group is 77.41%.

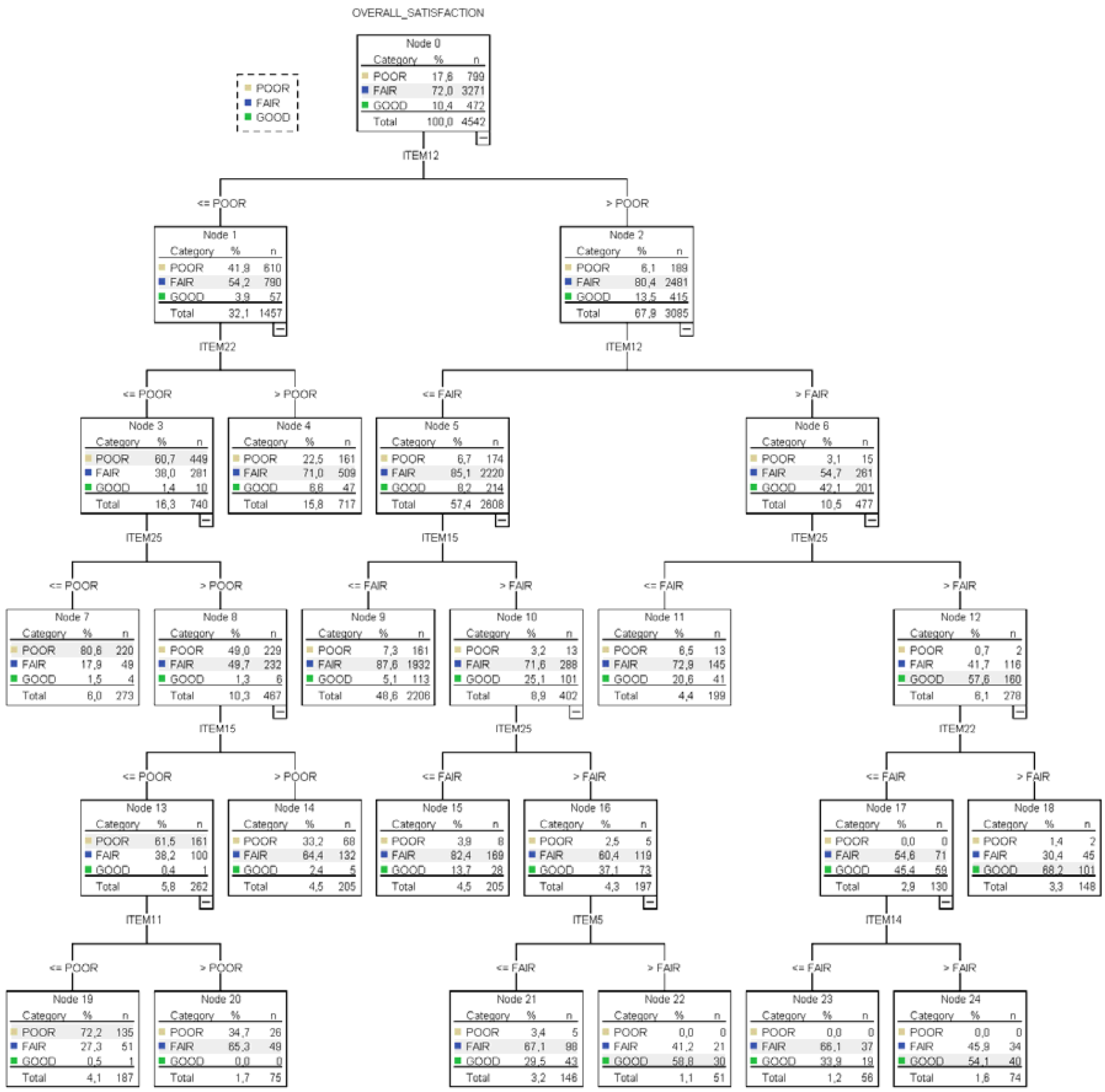


Figure 45. CART for users classified according to the frequency of the trip (“Daily”)

For “Weekly” group, a small tree was created (with accuracy rate of 77.66%) with only 3 terminal nodes: one node predicts a POOR overall satisfaction, while the other two a FAIR overall satisfaction. This tree only splits according to the Item15 “Regularity of runs” and Item23 “Complaints”. When these two variables are Poor, the “Overall satisfaction” also will be POOR. The tree built for passengers occasionally travelling is different from the other two groups. When passengers judge Item25 “Courtesy and competence on board” as Poor, the probability of having a POOR “Overall satisfaction” increases; on the other hand, to achieve that users have a GOOD satisfaction with the service, not only “Courtesy and competence on board” must be higher than Poor, but also Item12 “Fare/service rate” and Item4 “Cleanliness of vehicles” should be Good. This tree also achieve a high performance (77.53% of accuracy).

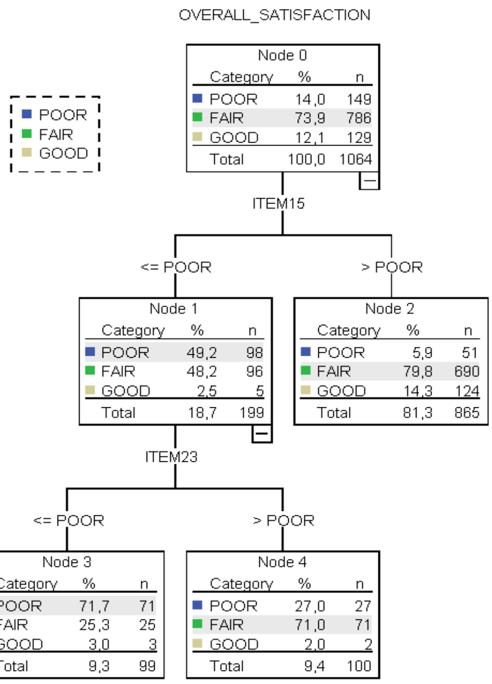


Figure 46. CART for users classified according to the frequency of the trip ("Weekly")

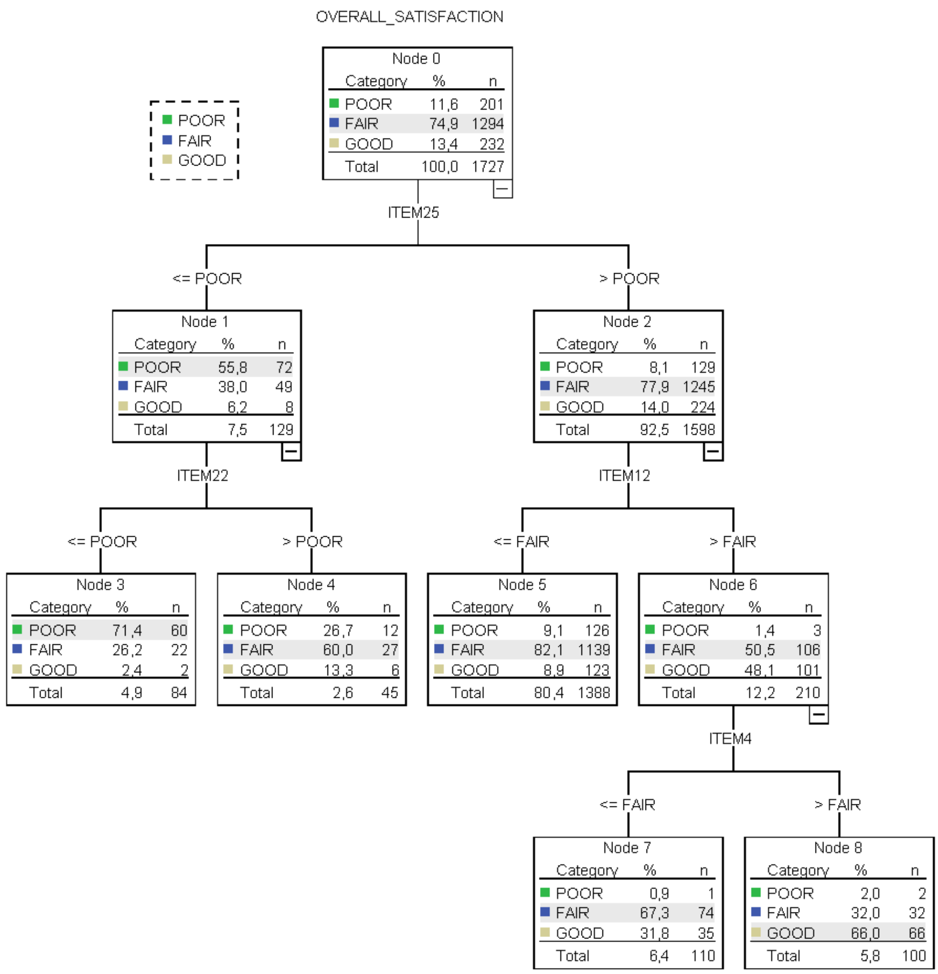


Figure 47. CART for users classified according to the frequency of the trip ("Occasionally")

5.3.2.2.3. Time of the trip market segments.

Finally, four different trees were built according to the time of the trip (Figures 48 to 51): “Off-peak hour”, “Morning peak hour”, “Afternoon peak hour”, and “Evening peak hour”. The precision rates of the trees have high and similar values,76.22% for the first group. 76.19% for the second, and 75.90% and 77.96% for the last two segments.

For “Off-peak hour” and “Afternoon peak hour” group, the splitter is Item12 “Fare/service rate”. For “Morning peak hour” group the splitter is Item10 “Air conditioning on board”, and for “Evening peak hour” group is Item25 “Courtesy and competence on board”.

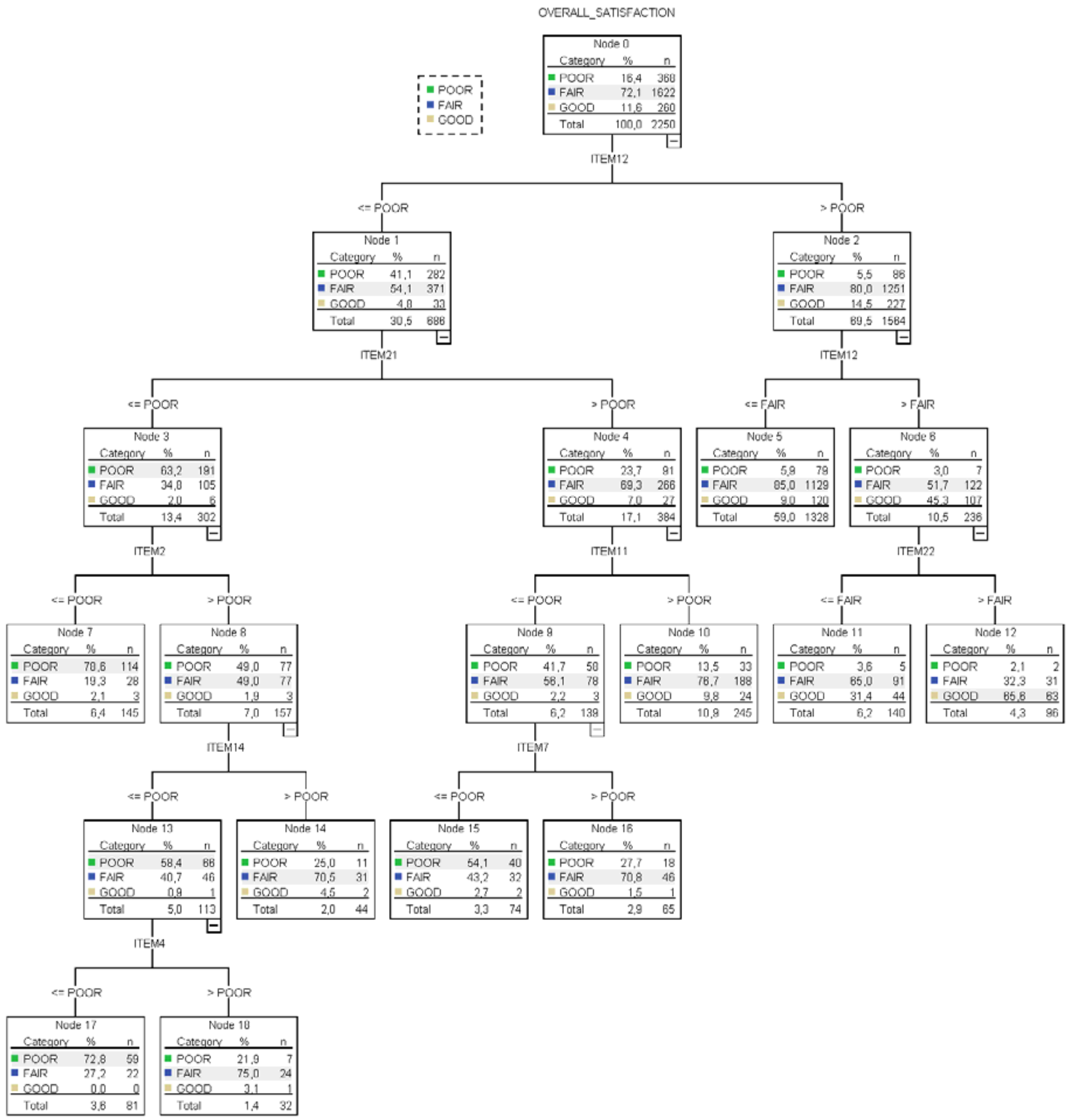


Figure 48. CART for users classified according to the time of the trip (“Off-peak hour”)

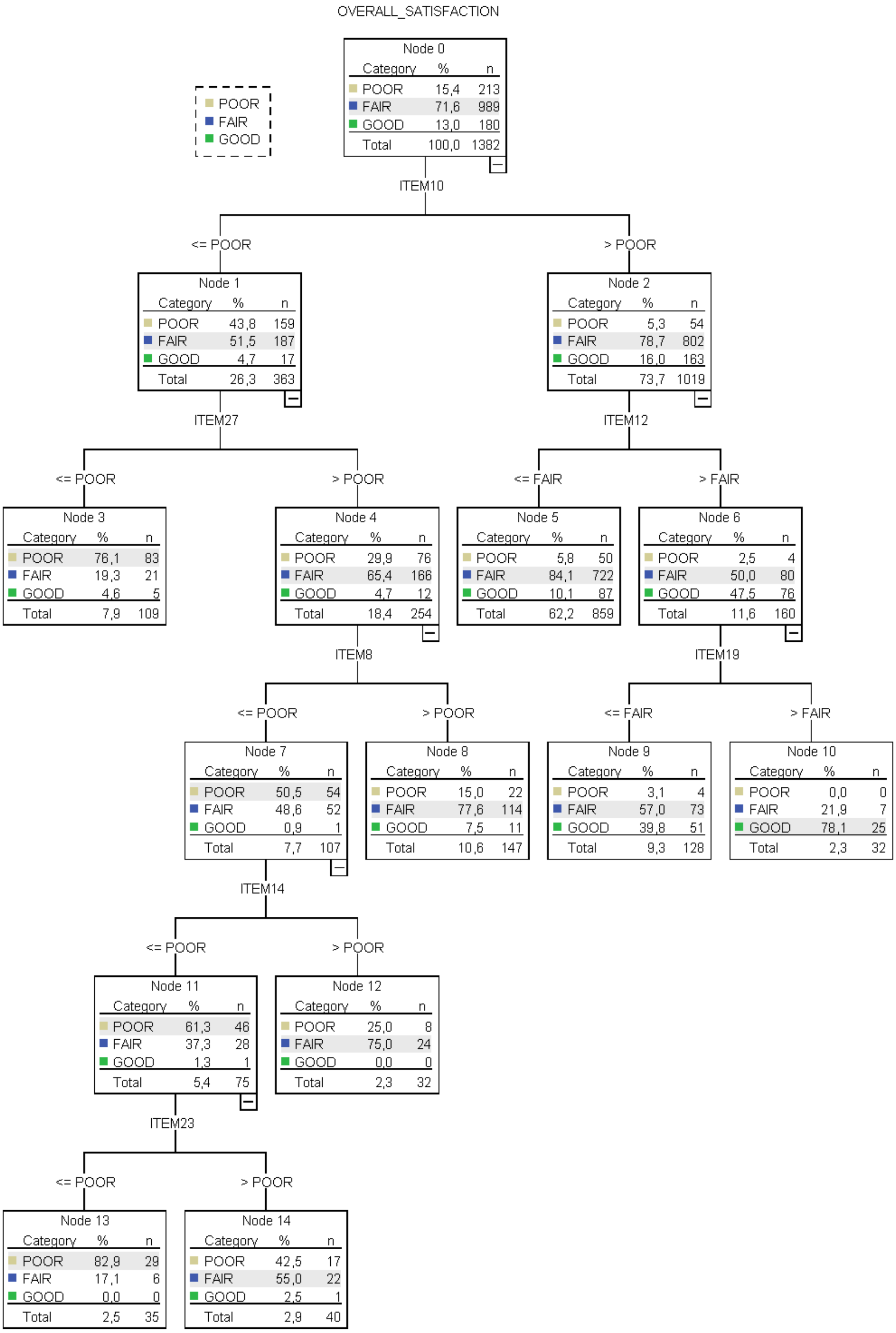


Figure 49. CART for users classified according to the time of the trip (“Morning peak hour”)

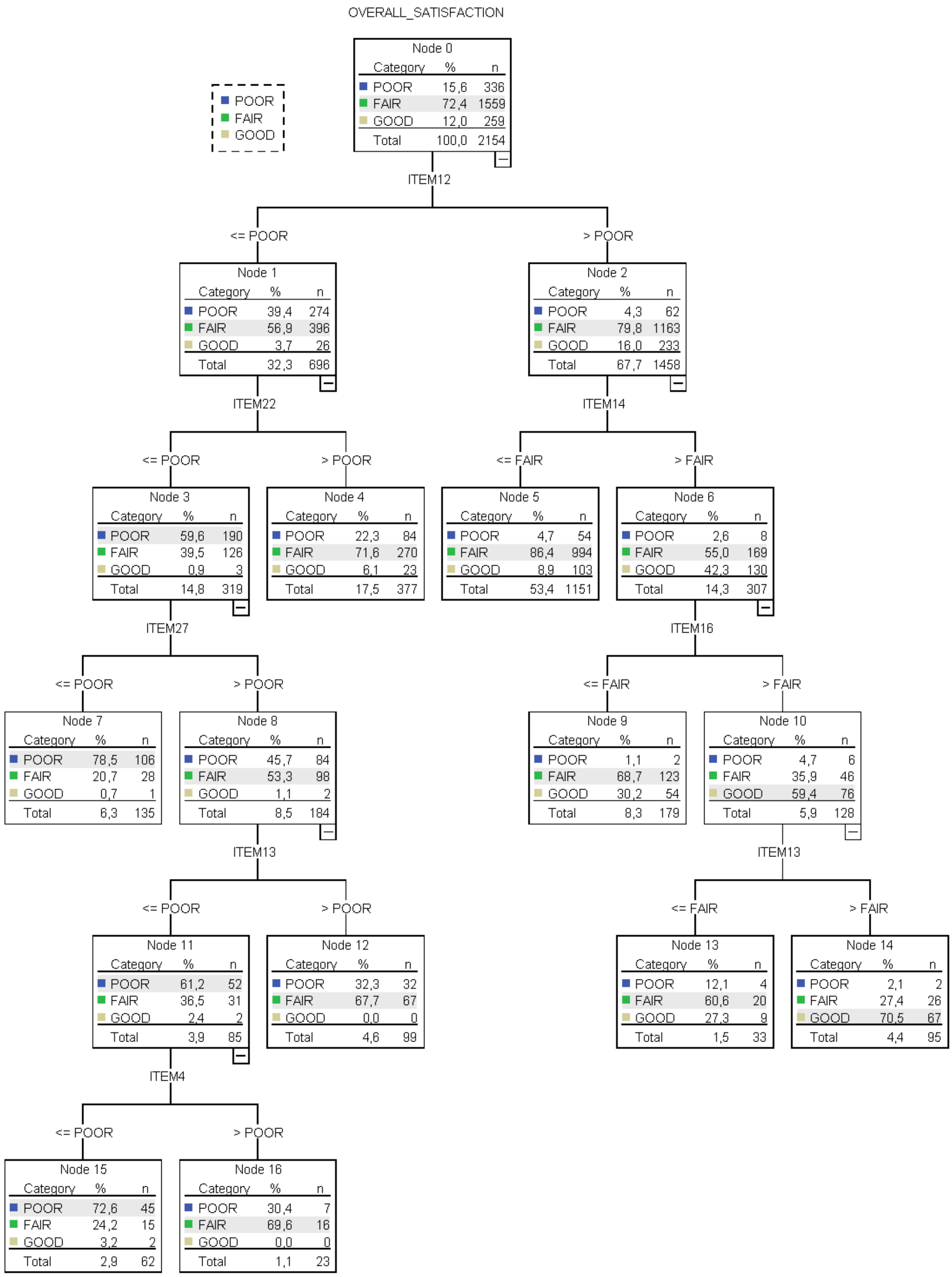


Figure 50. CART for users classified according to the time of the trip ("Afternoon peak hour")

The tree built for "Off peak hour" divides on the one side the passengers that has a Poor satisfaction with "Fare/service rate" and "Overall satisfaction" will not be GOOD, and on the other side those with a Fair or Good satisfaction with "Fare/service rate", and "Overall satisfaction" will be FAIR or higher. From the tree for "Morning Peak hour" it emerges that for achieving a GOOD overall satisfaction of the service in the "Morning peak hour" group, not only Item10 "Air conditioning on board" should not be Poor, but also Item12 "Fare/service rate" and Item19 "Bicycle transport on board" should have a Good satisfaction (78.1%). "Afternoon peak hour" tree, also divides at the left branch passengers that have a POOR or FAIR "Overall Satisfaction" with the service, and at the right branch, those with a FAIR or higher "Overall Satisfaction". The tree built for "Evening peak hour" group only created two terminal nodes; when "Courtesy and competence on board" is considered as Poor, also the "Overall satisfaction" will be POOR, with a probability of 64.8%.

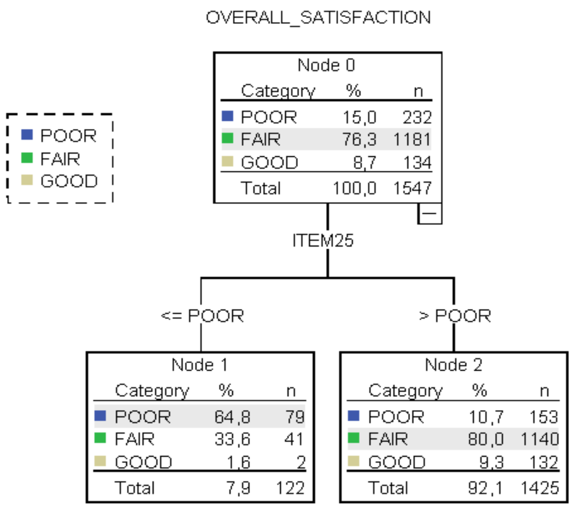


Figure 51. CART for users classified according to the time of the trip ("Evening peak hour")

5.3.3. Importance of the variables

Table 20 shows the normalized importance of the variables deduced from each of the models developed. This importance is extracted by the importance index (Kashani and Mohaymany, 2011). For simplicity's sake, the table only shows the three most important variables in each case.

For the tree contructed with the overall market, the factors deduced as most influential were the "Regularity of runs" and "Punctuality of the runs", which represent characteristics peculiar to a transit service.

By observing the most important attributes in the market segments identified by type of users, there are differencies among the three categories of users (Table 20). For "Commuter Workers" and also for "Others" the most important attribute is "Regularity of runs". For "Commuter students", however, the most important attribute is "Fare/service rate". This could be explained by considering that these two groups are the most frequent passengers and they spend more money for travelling. For the category "Others" the attribute "Frequency of the runs" is very important; it is not a relevant attribute for the other two groups, maybe because they know well the timetable of the service and they worry about other aspects of the service.

Concerning the trees built for different type of day, the most important variables are identified and are verified to be different among segments (Table 20). For “Working days”, many variables have a great impact in the prediction of the overall satisfaction, but the most important is “Regularity of the runs”, followed by “Punctuality of the runs” and “Courtesy of the personnel on board”; people travelling in the working days give more importance, as expected, to the aspects peculiar of a transit service. For the group “Days before holiday” and also for “Holidays”, few variables have importance. As an example, “Courtesy of the personnel at the stations”, “Courtesy of the personnel on board”, and “Safety” have a strong impact for “Days before Holiday” group, which includes people not travelling every day for whom more qualitative aspect could be important; for the group “Holidays” the unique very important variable is the “Fare/service rate”, maybe because people travelling in no working days are more demanding in terms of fare/service rate.

Regarding the Frequency of the trip, also high differencies were identified among the most important variables in the models. For users daily travelling, the most important variables are linked to both more qualitative and less qualitative service aspects (“Courtesy and competence on board”, “Regularity of runs”, “Information on board”) (Table 20). For users weekly travelling the most important variable is “Information at stations”, followed by “Windows and doors working” and “Regularity of runs”; for this kind of users the aspects linked to the information to users are obviously more relevant because they know the service less than the habitual users and need information for travelling. However, passengers occasionally travelling only focus on the “Courtesy and competence on board”, “Courtesy and competence on stations”, and “Fare/service rate”; this group of users prefers to be sure of having a good treatment from the personnel while travelling and they are less interested to other aspects of the service.

Finally, the factors retained as most important by users depending on the time of the trip are determined. “Courtesy and competence on board” is the most important variable for “Morning peak hour” and “Evening peak hour” groups, maybe because people, being scarcely awake in the morning and tired in the evening, need a kind treatment from the personnel. “Fare/service rate” has a high importance in the morning peak hour. In the off-peak hour “Regularity of runs” is the most important because in that period of the day the number of runs is not very high and people need a more regular service; finally, in the afternoon peak hour the “Frequency of runs” is in the first position.

Table 20. Importance for users of suburban lines and for users classified according to the four criteria (type of user; day of the trip; frequency of the trip; time of the trip)

Criteria of	*Category of User*	*Independent Variable*		*Normaliz*
SUBURBAN LINES (n. obs. 7,333; prec. rate 78.15%)		Item15	Regularity of Runs	100.0%
		Item14	Punctuality of Runs	93.7%
		Item4	Cleanliness of Vehicles	86.6%
TYPE OF USER (n. obs. 7,333)	COMMUTER WORKERS (n. obs. 2,371; prec. rate 78.87%)	Item15	Regularity of Runs	100.0%
		Item21	Information at Stations	93.7%
		Item22	Information on Board	93.4%
	COMMUTER STUDENTS (n. obs. 2,000; prec. rate 77.60%)	Item12	Fare/Service Ratio	100.0%
		Item25	Courtesy and Competence	84.7%
		Item11	Windows and Doors	78.6%
	OTHERS (n. obs. 2,962; prec. rate 77.28%)	Item15	Regularity of Runs	100.0%
		Item23	Complaints	81.6%
		Item13	Frequency of Runs	80.4%
DAY OF THE TRIP (n. obs. 7,333)	WEEKDAYS (n. obs. 6,207; prec. rate 76.83%)	Item15	Regularity of Runs	100.0%
		Item14	Punctuality of Runs	96.3%
		Item25	Courtesy and Competence	85.9%
	DAYS BEFORE HOLIDAY (n. obs. 600; prec. rate 77.50%)	Item27	Courtesy and Competence	100.0%
		Item25	Courtesy and Competence	96.6%
		Item1	Travel Safety	80.2%
	HOLIDAYS (n. obs. 526; prec. rate 76.43%)	Item12	Fare/Service Ratio	100.0%
		Item3	Personal Security at Station	66.5%
		Item13	Frequency of Runs	60.1%
FREQUENCY OF THE TRIP (n. obs. 7,333)	DAILY (n. obs. 4,542; prec. rate 77.41%)	Item25	Courtesy and Competence	100.0%
		Item15	Regularity of Runs	99.3%
		Item22	Information on Board	94.5%
	WEEKLY (n. obs. 1,064; prec. rate 77.63%)	Item21	Information at Stations	100.0%
		Item11	Windows and Doors	92.4%
		Item15	Regularity of Runs	87.2%
	OCCASIONALY (n. obs. 1,727; prec. rate 77.53%)	Item25	Courtesy and Competence	100.0%
		Item12	Fare/Service Ratio	93.2%
		Item27	Courtesy and Competence	77.9%
TIME OF THE TRIP (n. obs. 7,333)	OFF-PEAK HOUR (n. obs. 2,250; prec. rate 76.22%)	Item15	Regularity of Runs	100.0%
		Item4	Cleanliness of Vehicles	99.8%
		Item11	Windows and Doors	99.6%
	MORNING PEAK HOUR (n. obs. 1,382; prec. rate 76.19%)	Item25	Courtesy and Competence	100.0%
		Item12	Fare/Service Ratio	99.1%
		Item27	Courtesy and Competence	94.4%
	AFTERNOON PEAK HOUR (n. obs. 2,154; prec. rate 75.90%)	Item13	Frequency of Runs	100.0%
		Item14	Punctuality of Runs	95.7%
		Item15	Regularity of Runs	95.3%
	EVENING PEAK HOUR (n. obs. 1,547; prec. rate 77.96%)	Item25	Courtesy and Competence	100.0%
		Item27	Courtesy and Competence	81.7%

5.4. General Discussion

CART methodology has been applied in this Ph.D. thesis in order to analyze service quality in public transportation. This new approach allows predicting the level of overall service quality of a public transportation as well as identifying the key factors influencing this quality. Another interesting result of this methodology are the "If-then" rules generated by the different models, which provide to managers and PT operators very practical and useful information on which to base effective decision-making to promote the use of PT.

The accuracy rates of the trees generated in the different experimental contexts (models built for the bus public transport with and without market segmentation and models built for the rail public transport according to market segmentation) obtained high values, ranging from 59.72% to 79.12% in all of them.

On the other hand, the importance of the variables was successfully identified in each of the models built and it was verified that the importance rates stated by the passengers were different to those derived by the models, not showing the stated importance rates declared by the users the real influence of these variables in their global evaluation about the service quality. In general, when you ask passengers the importance of the variables describing the service, they consider that all the attributes are highly important for the performance of the service. This is one of the serious drawbacks encountered when studying the importance of variables based on the stated opinions of passengers.

Moreover, the "If-then" rules extracted from the models had also satisfactory values of Confidence and Support (rules analyzed in the Pre-Evaluation and Post-Evaluation bus public transport). For the Confidence rate, the values ranged between 45.8% and 76.2%, being these values the more extreme ones. The decision rules generated in the Post-Evaluation model achieved better rules' precisions, higher than 60% almost in all the cases. The Support rates were higher than 1% of the sample in all the cases.

By observing the results obtained in the analysis carried out for the bus public service with the data of 2007, they supported the main objectives of the doctoral thesis: CART model evaluated effectively service quality in the metropolitan bus transit service and it also identified the key factors affecting this quality. In addition, the hypothesis regarding that passengers make a different evaluation about the service quality before and after making them reflect on the attributes defining the service was also verified.

With the analysis of the two models it was found that passengers' perceived quality of service was practically limited to Frequency, Speed and Punctuality in the preliminary evaluation. Once they were made to reflect on other aspects that define the service, however, of other quality-related attributes gained in importance, such as Proximity to the bus stop, Safety on board and the Fare, while the impact of the previous three attributes diminished (Frequency, Speed and Punctuality).

These measures can be used to enhance service attributes to which passengers are unconsciously more sensitive in their preliminary evaluation, or to enhance the service attributes that are more important in the second evaluation. This last approach, however, would not make sense unless it is accompanied at the same time by a publicity campaign highlighting the service functions that passengers did not pay attention to at first.

Concerning to the research developed through market segments the differences in the key factors affecting service quality between the overall market and the various market segments were identified. Therefore, it was verified twice (one for the bus public transport context and the other one for the rail public transport context) that the main attributes leading to service quality tend to change, depending on the market segment under study. Therefore, when analyzing SQ, it is advisable to take different groups of users into consideration, so transport planners can direct their efforts more accurately to the group of users whose loyalty they seek by attending to their preferences and needs.

Normally the segmentation includes frequent and sporadic users, or users are grouped by sex, age or minimum income. In this study, in addition to previous segmentations, it was also used less common groups (i.e. travel reason and type of ticket in the bus public service, and type of user, type of day and time of the trip in the rail public service) that also reveal interesting results. These analysis also supported the main objectives of the thesis.

For the bus public experimental context Punctuality was the most important SQ attribute at the level of the overall market and in most market segments, as opposed to Fare, which was the least relevant attribute. Accesibility is shown to be highly important, mainly for passengers who use the senior citizen pass, whereas it is not really important in most of the other segments.

Another aspect that has been brought up again in this analysis is the drawback of using stated importance methods for identifying the importance of each attribute. Some of the method's drawbacks (it increases the length of the survey; yields insufficient differentiation among mean importance ratings; and attributes may be rated as important even though they have little influence on SQ) have been pointed out before in the literature (Weinstein, 2000). To these it could be added two more drawbacks indicated in this research: the possibility of attributes that are important for passengers and yet are not identified as such in the survey (i.e. Information in this study); and the impossibility of identifying differences between market segments on the basis of users' direct answers with regards to importance.

The decision trees built provided good predictions, with accuracy values above 63.65%. In some market segments, even higher accuracy was attained (71.98% for Men, 78.98 % for Elderly, 69.77% for Sporadic, 69.58% for Other travel reason and 79.12% for Senior Citizen pass). This demonstrated that such segmentations could lead to sample homogeneity, as well as better results from the models.

According to the results obtained in the rail public service experimental context, the accuracy rates obtained were also high among the market segments, ranging from about 76.00% to 79.00% in all of them. In view of the previous research carried out in the bus public transport, the demographic variables and travel habits were not used as predictors of the models given that they were not identified as key factors in any of the previous models. Then, only the SQ attributes describing the service were considered in this part of the research as independent variables.

Different structures of the trees as well as different key factors were identified among the different groups of users. The variable that split the root node in most of the cases was Fare/service ratio (for the overall market and most of the segments). Some specific aspects of the service, such as Regularity or Punctuality of Runs, are mostly considered by the habitual travellers, while the most qualitative aspects, such as Courtesy and competence on board, are preferred by others. This is a very interesting findings, as very interesting are all the suggestion given by the analysis of the results.

For example, commuter workers and students considered the aspect linked to the Fare/service rate as important, maybe because they travel frequently. People travelling in working days give importance to specific aspects of a transit service, such as Frequency, Regularity and Punctuality of the runs, while people travelling in no working days focus their attention to aspects such as Courtesy of the personnel or Safety. People daily travelling are interested to many service aspects (peculiar of the service and more qualitative), while users weekly travelling focus their attention to the Information, because they know the service less and need information for travelling. For passengers occasionally travelling and those travelling in the morning and evening peak hours it is very important to receive a kind treatment from the personnel. People travelling in the off-peak hour consider Regularity of runs as the most important aspect because in that period of the day the number of runs is not very high and people need a more regular service.

We can conclude that this research may be useful to public transport planners in a number of ways:

- First, to predict the level of quality is being provided

- Second, to extract descriptive "If-Then" rules which can explain the interaction between variables and provide practical and valuable information.

- Third, to clarify the factors that have a notable impact on service quality, either overall or by user segments.

- Fourth, as the factors are not the same for all passengers, each market segment will require different incentives (i.e. personalized marketing).

The ability to differentiate between key service quality factors in the various market segments will enable transport planners to decide which users they want to engage and plan their loyalty programs accordingly.

CHAPTER 6

CONCLUSIONS AND FUTURE RESEARCH

Chapter 6

CONCLUSIONS AND FUTURE RESEARCH

6.1. Conclusions

In this chapter the major conclusions of this Ph.D. thesis are pointed out.

In general, Chapter 2 summarizes the existing literature in the analysis and modeling of service quality in public transportation. The complexity of the concept and the different challenges that should be considered for its measure have been highlighted in this literature review, such as knowing the relationship between SQ and satisfaction; identifying the most relevant attributes that affect SQ; dealing with subjective, qualitative and fuzzy data; using or not using objective data, the limitations of customers satisfaction surveys, the heterogeneity of passengers'opinions; etc.

Several methodological approaches have been used. While practitioners, transport operators and governments focus their analysis in disaggregated models, such as quadrant analysis, for evaluating the service provided and for setting priorities for service improvements among a long list of service attributes, the preferred techniques for analyzing service quality by researchers are those reaching a global indicator that could be used to compare different services and their development over time. For them, the most used techniques are those based on disconfirmation theory. However, in recent years, both approaches (aggregated and disaggregated) are being used to complement each other. Each method has its own specifications and limitations, and it is not an easy issue identify which is the best statistical method for analyzing service quality.

The heterogeneity present in users' perceptions is one of the main concerns of researchers. Stratifying the sample of users in most homogeneous subsamples with similar perceptions is one of the most applied and effective solutions for managing this heterogeneity. However, other mehtods can be considered in order to solve this problem, such us the HCSI proposed by Eboli and Mazzulla (2009) or the mixed logit models, which have the ability to handle this heterogeneity.

On the other hand, derived importance methods, which determinate the importance of the attributes by statistically testing the strength of the relationship of individual attributes with overall SQ, are preferred by researches because of their numerous advantages, however, asking customers to rate each attribute on an importance scale is still the method mostly used by researchers and operating companies, because of their simplicity and also because of they do not need to establish different assuptioms or underlying relationships among variables, which on most occasions are violated.

Decision Trees, and particularly CART methodology, has been proposed in this Ph.D. thesis for analyzing SQ in public transportation due to its numerous advantages over other parametric methodologies. The main benefits of this technique are: it does not need to establish a functional form, it can handle large data bases and the complex interactions and patterns among data are easily identified, large quantity of explanatory variables are able to be used and the importance of these variables in the model are easily found, the outcomes of the model are displayed in understanding graphics, useful "If-Then" decision rules can be extracted, the mullticollinearity and outliers do not represent a problem for this technique, etc.

SQ data were obtained from various CSS devoloped in two different public transport services: the Granada metropolitan bus service and a suburban rail service in the north of Italy. The first experimental context was also split in two cases of study: one for the data collected in 2007, and the other one for the data collected in the period of time 2008 - 2011. Several DTs were built using the CART algortihm and the main conclusions of these analysis were:

- The accuracy indicator was used to evaluate the gobal performance of the DTs, and the confidence and support rates were used to define the quality of the decision rules. By observing the results of the different models, they showed good accuracy values, in few cases similar to other studies with similar purposes (Wong and Chung, 2007), but in general, most part of the models achieved higher accuracies rates than previous research in this field. According to the confidence and support of the extracted decision rules, they obtained satisfactory values in almost all of them, providing useful information about the interaction of variables for practitioners and operators.

- The variable importance index identified the most important variables in evaluating SQ in all the study cases of this Ph.D. thesis. Identifying the most relevant variables influencing the overall service quality perceived by users and determining the weight of them in the model was one of the specific objectives of this research. Its results were compared with those stated by the users in the CSS, and the drawbacks of using stated importance methods for identifying the importance of each attribute was verified. Some of the method's drawbacks (it increases the length of the survey; yields insufficient differentiation among mean importance ratings; and attributes may be rated as important even though they have little influence on SQ) have been pointed out before in the literature (Weinstein, 2000). To these it could be added two more drawbacks indicated in this study: the possibility of attributes that are important for passengers and yet are not identified as such in the survey and the impossibility of identifying differences between market segments on the basis of users' direct answers with regards to importance. This is the reason why deriving the importance of the attributes should receive more attention by practitioners and researchers in order to understand what really influence passengers overall perceptions about the service quality. Then, the existing problematic of stated importance technique for determining the importance of the variables over the overall service quality has been inquired in detail in this research complying another specific objective of this thesis.

- The hypothesis regarding to passengers change their evaluation about the service quality before and after reflecting on the attributes that describe the service was also confirmed in the first study case, and the variables that played the most important role in their evaluations were identified. While in their previous evaluation their overall perception about the quality is lower and it is unconciously limited to few attributes of the service (Frequency, Speed and Punctuality), when they are made to reflect about the different attributes describing the service, other quality-related attributes gained in importance, such as Proximity to the bus stop, Safety on board and the Fare and the impact of the previous attributes diminish.

- The heterogeneity present in passengers' opinions could be reduced by stratifying the sample of users in more homogeneous groups. By observing the results of the second and third study cases it was proved that the variables most influencing passengers overall evaluation were different among market segments. Therefore, when analyzing SQ, it is advisable to take different groups of users into consideration so transport planners can direct their efforts more accurately to the group of users whose loyalty they seek by attending to their preferences and needs.

- It should be pointed out that due to DTs permit to use large amount of independent variables of diverse nature on their models, and the most important are easily identified, different varibles of different nature (socioeconomic variables, travel habits and SQ attributes) were used as predictor in the models built for the bus service experimental context. However, the variables that reach the higher influence over the overall service quality were those related with the SQ attributes. For this reason, in the rail service experimental context only the attributes describing SQ characteristics were used as predictors and sucessfully results were achieved.

- Finally, it can be concluded that DTs is an adequate technique for analyzing SQ in public transportation. This technique can be used not only for predicting and classifying SQ, but also for extracting useful decision rules, and deriving the importance of the variables in the model. So the main hypothesis of this research work is confirmed: "Decision Trees is an appropiate methodology for analyzing service quality in public transportation".

- The results of this analysis should not be generalized to other type of public transport services (such as urban public tansport services or even metropolitan or suburban public transport services using different modes of transport than the ones analyzed here, such a metro system) because the performance characteristics and passengers' requirements differ widely among transit services. The policies for SQ in public transportation can only succeed if they apply specific measures focusing on the characteristics of the type of service under study and their specific needs; a generic framework of action is not recommended. However, the used of this methodology can be applied for analyzing service quality in any type and context of public transportation.

- In addition, it can be highlighted that the data used in this research work come from various CSS that were non-research oriented surveys. A rather simple statistical frequency analysis was the main target. However, the application of more advanced modelling techniques proves that this kind of data can be used to reveal very interesting details for managers and public transport operators and it could increase the collaboration between researchers and the industry.

- However, some limitations have been found carrying out this research that should be pointed out. Usually, the data collected in the CSS show imbalanced data of the passengers' perceptions. Subsequently, the classifier produces biased results, with high predictive accuracy over the over-represented classes, but a low predictive accuracy over the under-represented classes.

- So, to sum up, Decision Trees have a great interest for public transport managers. While other more popular statistical techniques used for analyze service quality in public transportation need that statisticians or professional users interpret the results, Decision Trees provide practical information for public transport managers because of their simplicity, the easiness of understanding, the possibility of formulating decision rules, the ability of extracting the importance of each variables, etc.

6.2. Future research work

From the elaboration of this Ph.D. thesis and analyzing the main conclusions obtained from this research work, some studies are planned to be developed in the near future. Thus, in terms of future work, it would be interesting to performe the following research lines:

- Due to CART algorithm produce binary splits, and sometimes the influence of specific categories of the independent variables are imposible to be analyzed, another future work planned it to apply other DT algorithm (for example the C5.0 algortihm) that permits obtaine this information. Comparison between the tree models generated by both algortihms (CART and C5.0) could be carried out in order to identify the most important attributes and more powerful decision rules when the variables inducing them are coincidential in both algorithms. Moreover, different information could be revealed by both methods, which instead of exclude each other, they could be complementary and appropiate for a full understanding of the phenomenon analyzed.

- Model SQ using Artificial Neural Networks and compare the results with the ones extracted by Decision Trees. Artificial Neural Networks is another data mining technique widely used in transportation problems, such as choice behaviour. They are information-processing structures designed to model the behavior of human neurons (Xie et al. 2003). They usually reach high accuracy predictions but the results interpretation is more difficult than for Decision Trees.

- Predict SQ by Regresion trees, predicting a value of the overall quality and not a classification of this quality. When transport managers want to prioritize the strategies that are going to develope on the service, initially they search by indications on a large scale (if the service is Poor, Fair or Good) about where are the main weakness of the service. Once they reach providing high quality services, Regression trees seem to offer more detail information regarding to the evaluation about the service quality.

- Reduce the heterogeneity of passengers' opinions by using cluster analysis for stratifying the sample of users. The analysis cluster is a data mining technique that permits identify homogeneous groups of users that share a set of characteristics. For example, a cluster of users could be represented by middle age women that do not have avaiable a private vehicle and they are commuter workers. From the extracted clusters identified in the dataset, different DTs will be built and the results will be compared with the ones obtained in the model generated with the whole dataset, and models generated with the traditional segmentation (e.g. gender, age, etc), in order to find out if using a clustering technique for stratifying the sample affect the performance of the trees (higher accuracies rates are achieved) and different key factors over the overall service quality evaluation are identified among clusters.

- Treat the class predictive accuracy problems of the decision trees produced by the imbalanced dataset. This problem has been handled by different authors in different fields (Ling and Li, 1998; Kubat and Matwin, 1997; Kubat et al., 1998; Montella et al. 2012; Riddle et al. 1994)

- DTs methodology has been applied in this research work into two different experimental contexts (a metropolitan bus service and a suburban rail service) regarding to the type of service (bus vs. rail) and the country were it is performed (Spain vs. Italy). Future research work could be focus on analyzing other types of public transport services, and collaborate with other research institutions of different countries in order to compare results among them.

CHAPTER 7

REFERENCES

Chapter 7

REFERENCES

Abalo J, Varela J, Manzano V. (2007) Importance values for importance-performance analysis: a formula for spreading out values derived from preference rankings. Journal of Business Research, 60(2), 115–121.

Abdlla G, Mohamed AR, Mekawy MA. (2007) Managing tourists' needs and expectations: An empirical analysis of the egyptian airline sector. Tourism, 55(3), 277-296.

Aksoy S, Atilgan E, Akinci S. (2003) Airline services marketing by domestic and foreign firms: differences from the customers' viewpoint, Journal of Air Transport Management, 9, 343–351.

Andreassen TW. (1995) (Dis)satisfaction with public services: The case of public transportation. Journal of Services Marketing 9: 30–41.

Arentze TA, Hofman F, van Mourik H, Timmermans HJP, Wets G. (2000) Using decision tree induction systems for modeling space–time behavior. Geographical Analysis 32, 330–350.

Arantze TA, Timmermans HJP. (2007) Parametric action decision trees: Incorporating continuous attribute variables into rule-based models of discrete choice. Transportation Research Part B, 41, 772-783

Awasthi A, Chauhan SS, Omrani H, Panahi A. (2011) A hybrid approach based on SERVQUAL and fuzzy TOPSIS for evaluating transportation service quality. Computers and Industrial Engineering, 61(3), 637-646.

Babakus E, Boller GW. (1992) An Empirical Assessment of the SERVQUAL Scale. Journal of Business Research, 24, 253-268.

Bollen KA. (1989) Structural equations with latent variables. New York: Wiley.

Brady MK, Cronin Jr JJ. (2001) Some new thoughts on conceptualizing perceived service quality: a hierarchical approach. Journal of Marketing, 65(3), 34-49.

Breiman L, Friedman JH, Olshen RA, Stone CJ. (1984). Classification and Regression Trees. Wadsworth, Belmont (CA).

Brown TJ, Churchill Jr GA, Peter JP. (1993) Improving the Measurement of Service Quality. Journal of Retailing, 69(1), 127–139.

Burns AC. (1986) Generating marketing strategy priorities based on relative competitive position. Journal of Customer Marketing, 13(4), 49–56.

Carman J. (1990) Consumer perceptions of service quality: an assessment of the SERVQUAL dimensions. Journal of Retailing, 66, 33-55.

Caro LM, Garcia JAM. (2007) Measuring perceived service quality in urgent transport service. Journal of Retailing and Consumer Services, 14(1), 60-72.

Carrillat FA, Jaramillo F, Mulki JP. (2007) The validity of the SERVQUAL and SERVPERF scales A meta-analytic view of 17 years of research across five continents. International Journal of Service Industry Management, 18(5), 472-490.

Carter DN, Lomax TJ. (1992) Development and application of performance measures for rural public transportation operators. Transportation Research Record, 1338, 28–36.

Catillo JM, Benitez FG. (2012) Determining a public transport satisfactionindex from user surveys. Transportmetrica, 1-29, iFirst

Cavana RY, Corbett LM, Lo YL. (2007) Developing zones of tolerance for managing passenger rail service quality. International Journal of Quality & Reliability Management, 24(1), 7–31.

Chan KY, Loh WY. (2004) LOTUS: An algorithm for building accurate and comprehensible logistic regression trees. Journal of Computational and Graphical Statistics, 13, 826–852.

Chang H, Liang G, Chu C, Chou C. (2012) Prioritizing service attributes for improvement using fuzzy zone of tolerance. International Journal of Innovative Computing, Information and Control, 8(1A), 75-89

Chang LY, Wang HW. (2006) Analisys of traffic injury severity: An application of non-parametric classification tree techniques. Accident Analysis and Prevention 38, 1019–1027

Chang YH, Yeh CH. (2002) A survey analysis of service quality for domestic airlines. European Journal of Operational Research, 139, 166-177.

Chau VS, Kao YY. (2009) Bridge over troubled water or long and winding road? Gap-5 in airline service quality performance measures. Managing Service Quality, 19(1), 106-134.

Chen C. (2008) Investigating structural relationships between service quality, perceived value, satisfaction, and behavioral intentions for air passengers: Evidence from Taiwan. Transportation Research Part A, 42(4), 709-717.

Chen FY, Chang YH. (2005) Examining airline service quality from a process perspective. Journal of Air Transport Management, 11, 79–87.

Cheng J-H, Chen F-Y, Chang Y-H. (2008) Airline relationship quality: an examination of Taiwanese passengers. Tourism Research, 29, 487-499.

Choi KS, Cho WH, Lee S, Lee H, Kim C. (2004) The relationships among quality, value, satisfaction and behavioral intention in health care provider choice: a South Korean study. J. Bus. Res, 57(8), 913–921.

Chou J, Kim C, Kuo Y, Ou N. (2011a) Deploying effective service strategy in the operations stage of high-speed rail. Transportation Research Part E, 47(4), 507-519.

Chou CC, Liu LJ, Huang SF, Yih JM, Han TC. (2011b) An evaluation of airline service quality using the fuzzy weighted SERVQUAL method. Applied Soft Computing, 11, 2117-2128.

Chou JS, Kim C. (2009) A structural equation analysis of the QSL relationship with passenger riding experience on high speed rail: An empirical study of Taiwan and Korea Expert Systems with Applications, 36, 6945–6955.

Christopher MK, Stuart D, Foote PJ. (1999) Structuring and assessing transit management response to customer satisfaction surveys. Transportation Research Record, 1669, 99-108.

Churchill GA. (1979) A paradigm for developing better measure of marketing constructs. Journal of Marketing Research, 16, 64–73.

Cronin Jr JJ, Brady MK, Hult GTM. (2000) Assessing the effect of quality, value and customer satisfaction on consumer behavioral intention in service environment. Journal of Retailing, 76(2), 193-218.

Cronin J, Taylor S. (1992) Measuring service quality: a reexamination and extension. Journal of Marketing, 56, 55-68.

Dabholkar PA, Shepherd CD, Thorpe DI. (2000) A comprehensive framework for service quality: an investigation of critical conceptual and measurement issues through a longitudinal study. J. Retail., 76(2), 139–173.

Dabholkar PA, Thorp DI, Rentz JO. (1996) A measure of service quality for retail stores: scale development and validation. J. Acad. Mark. Sci., 24(1), 3–16.

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

Department for Transport (2005) Bus Quality Indicators: England. Available at: http://www.dft.gov.uk/stellent/groups/dft_transstats/documents/page/dft_transstats_041049.hcsp).

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

De Oña J, Lopez G, Abellán J. (2013) Extracting decision rules from police accident reports through decision trees. Accident Analysis and Prevention, 50, 1151-1160

De Oña R, Eboli L, Mazzulla G. (2012) Key factors affecting rail service quality. a decision tree approach. XIX Conference SIDT, Padua, 18-19 October, 2012.

Diana M. (2012) Measuring the satisfaction of multimodal travelers for local transit services in different urban contexts. Transportation Research Part A, 46(1), 1-11.

Drea JT, Hanna JB. (2000) Niche marketing in intrastate passenger rail transportation. Transportation Journal, 39(3), 33-43.

Eboli L, Mazzulla G. (2007) Service quality attributes affecting customer satisfaction for bus transit. Journal of Public Transportation, 10(3), 21-34.

Eboli L, Mazzulla G. (2008) A Stated Preference Experiment for Measuring Service Quality in Public Transport. Transportation Planning and Technology, 31(5), 509-523.

Eboli L, Mazzulla G. (2009) A new customer satisfaction index for evaluating transit service quality. Journal of Public Transportation, 12(3), 21-37.

Eboli L, Mazzulla G. (2010) How to capture the passengers' point of view on a transit service through rating and choice options. Transport Review, 30(4), 435-450.

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, 172–181.

Eboli L, Mazzulla G. (2012) Performance indicators for an objective measure of public transport service quality. European Transport, 51.

Esposito F, Malerba D, Semeraro G. (1997) A Comparative Analysis of Methods for Pruning Decision Trees. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19 (5), 476–491.

Fernandes E, Pacheco RR. (2010) A quality approach to airport management. Quality and Quantity, 44(3), 551-564.

Fielding GJ. (1992) Transit performance evaluation in the USA. Transportation Research, 26A(6), 483–491.

Fielding GJ, Babitsky TJ, Brenner ME. (1985) Performance evaluation for bus transit. Transportation Research, 19A(1), 73–82.

Figler SA, Sriraj PS, Welch EW, Yavuz N. (2011) Customer loyalty and Chicago, Illinois, transit authority buses. Transportation Research Record, 2216, 148-156.

Foote PJ, Stuart DG. (1998) Customer satisfaction contrasts express versus local bus service in Chicago's North Corridor. Transportation Research Record, 1618, 143–152.

Foote PJ, Stuart DG, Elmore-Yalch R. (2001) Exploring customer loyalty as a transit performance measure. Transportation Research Record, 1753, 93–101.

Forgas S, Moliner MA, Sánchez J, Palau R. (2010) Antecedents of airline passenger loyalty: Low-cost versus traditional airlines. Journal of Air Transport Management, 16(4), 229-233.

Fornell C, Johnson MD, Anderson EW, Cha J, Bryant BE. (1996) The American Customer Satisfaction Index: Nature, purpose and findings. Journal of Marketing, 60(4), 7-18.

Fowler Jr FJ. (1995) Improving Survey Question: Design and Evaluation. Applied Social Research Methods Series, Vol. 38 (Newbury Park, CA: Sage).

Friman M, Gärling T. (2001) Frequency of negative critical incidents and satisfaction with public transport services, II. Journal of Retailing and Consumer Services, 8(2), 105-114.

Friman M. (2004) Implementing quality improvements in public transport. Journal of Public Transportation, 7(4), 49–65.

Friman M, Edvardsson B, Garling T. (2001) Frequency of negative critical incidents and satisfaction with public transport services, I. Journal of Retailing and Consumer Service, 8(2), 95-104.

Ganesan-Lim C, Russell-Bennett R, Dagger T. (2008) The impact of service contact type and demographic characteristics on service quality perceptions. Journal of Services Marketing, 22(7), 550–561.

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

Gatta V, Marcucci E. (2007) Quality and Public Transport Service Contracts. European Transport, 36, 92-106.

Gilbert D, Wong RKC. (2003). Passenger expectations and airline service: a Hong Kong based study. Tourism Management, 24, 519–532.

Glascock J. (1997) Research on customer requirements for transit service design and delivery. Transportation Research Record, 1604, 121–127.

Golob TF. (2003) Structural equation modeling for travel behavior research. Transportation Research Part B, 37(1), 1-25.

González MEA, Comesaña LR, Brea JAF. (2007) Assessing tourist behavioral intentions through perceived service quality and customer satisfaction, J. Bus. Res., 60(2), 153–160.

Gourdin K, Kloppenborg TJ. (1991) Identifying service gaps in commercial air travel: the first step toward quality improvement, Transportation Journal, 31(1), 22–30.

Grönroos C. (1982) Strategic Management and Marketing in Service Sector, Marketing Science Institute, Cambridge, MA.

Grönroos C. (1984) A service quality model and its marketing implications. European Journal of Marketing, 18(4), 36-44.

Grönroos C. (1988) Service Quality: The six criteria of good service quality, Review of Business. New York: St. John's University Press.

Hallmark SL, Guensler R, Gomunung F. (2002) Characterizing on-road variables that affect passenger vehicle modal operation. Transportation Research Part D, 7, 81-98

Hensher DA, Mulley C, Yahya N. (2010) Passenger experience with quality-enhanced bus service: The Tyne and Wear 'superoute' services. Transportation, 37(2), 239-256.

Hensher DA. (2000) Service quality as a package: What does it mean to heterogeneous consumers? Institute of Transport Studies, The University of Sydney, July.

Hensher DA. (2001) Service quality as a package: What does it mean to heterogeneous consumers? In: 9th World Conference on Transport Research, 22-27 July, Seoul.

Hensher DA, Daniels R. (1995) Productivity measurement in the urban bus sector. Transport Policy, 2(3), 179–194.

Hensher DA, Prioni P. (2002) A Service Quality Index for Area-wide Contract Performance Assessment. Journal of Transport Economics and Policy, 36(1), 9-113.

Hensher DA, Stopher P, Bullock P. (2003) Service quality - developing a service quality index in the provision of commercial bus contracts. Transportation Research Part A, 37(6), 499-517.

Hill N, Brierley G, MacDougall R. (2003) How to Measure Customer Satisfaction. Gower Publishing, Hampshire.

Hodgson P, Potter S. (2010) Refining light rapid transit typology: a UK perspective. Transportation Planning and Technology, 33 (4), 367-384

Hu K-C. (2010) Evaluating City Bus Service Based on Zone of Tolerance of Expectation and Normalized Importance. Transport Reviews, 30 (2), 195–217.

Hu K-C, Jen W. (2006) Passengers' Perceived Service Quality of City Buses in Taipei: Scale Development and Measurement. Transport Reviews, 26 (5), 645–662.

Huang C, Hsueh S. (2010) Customer behavior and decision making in the refurbishment industry - A data mining approach. Journal of Civil Engineering and Managing, 16(1), 75-84.

Huse C, Evangelho F. (2007) Investigating business traveller heterogeneity: Low-cost vs full-service airline users? Transportation Research Part E, 43, 259–268.

Jabnoun N, Khalifa A. (2005) A customized measure of service quality in the UAE. Managing Service Quality, 15(4), 374-388.

Jen W, Hu K-C. (2003) Application of perceived value model to identify factors affecting passenger's repurchase intentions on city bus: a case study of the Taipei metropolitan area. Transportation, 30, 307–327.

Jen W, Tu R, Lu T. (2011) Managing passenger behavioral intention: an integrated framework for service quality, satisfaction, perceived value, and switching barriers. Transportation, 38, 321–342.

Joewono TB, Kubota H. (2007a) Exploring negative experiences and user loyalty in paratransit. Transportation Research Record, 2034, 134-142.

Joewono TB, Kubota H. (2007b) User perceptions of private paratransit operation in Indonesia. Journal of Public Transportation, 10(4), 99-118.

Joewono TB, Kubota H. (2007c) User satisfaction with paratransit in competition with motorization in Indonesia: Anticipation of future implications. Transportation, 34(3), 337-354.

Joreskog KG. (1973) Analysis of covariance structures. In Multivariate Analysis-III, P. R. Krishnaiah, ed. New York: Academic Press, 263–285.

Karlaftis MG. (2004) Predicting Mode Choice through Multivariate Recursive Partitioning. Journal of Transportation Engineering, 130, 245-250

Karlaftis MG, Golias J, Papadimitriou E. (2001) Transit quality as an integrated traffic management strategy: Measuring perceived service. Journal of Public Transportation, 4(1).

Kashani AT, Mohaymany AS. (2011) Analysis of the traffic injury severity on two-lane, two-way rural roads based on classification tree models. Safety Science, 49, 1314 – 1320

Kass GV. (1980) An Exploratory Technique for Investigating Large Quantities of Categorical Data. Applied Statistics, 29,119-27.

Kiatcharoenpol T, Laosirihongthong T. (2006) Innovations in service strategy: An evaluation of quality in airline service operations by using SERVQUAL Model. In: IEEE international conference on management of innovation and technology, Singapore, China (pp. 748–752).

Kim YK, Kim YB, Lee YI. (2011) Perceived service quality for south korean domestic airlines. Total Quality Management and Business Excellence, 22(10), 1041-1056.

Kim YK, Lee HR. (2011) Customer satisfaction using low cost carriers. Tourism Management, 32(2), 235-243.

Koushki PA, Al-Saleh OI, Al-Lumaia M. (2003) On management's awareness of transit passenger needs. Transport Policy, 10, 17–26.

Kuhnert PM, Do KA, McClure R. (2000) Combining non-parametric models with logistic regression: an application to motor vehicle injury data. Computational Statistics & Data Analysis 34, 371-386

Kuo M-, Liang G-. (2011) Combining VIKOR with GRA techniques to evaluate service quality of airports under fuzzy environment. Expert Syst Appl. 38(3):1304-1312.

Kuo M, Wu J, Pei L. (2007) A soft computing method for selecting evaluation criteria of service quality. Applied Mathematics and Computation, 189(1), 241-254.

Kuo M. (2011) A novel interval-valued fuzzy MCDM method for improving airlines' service quality in chinese cross-strait airlines. Transportation Research Part E, 47(6), 1177-1193.

Lai W, Chen C. (2011) Behavioral intentions of public transit passengers-the roles of service quality, perceived value, satisfaction and involvement. Transport Policy, 18(2), 318-325.

Landrum H, Prybutok VR. (2004) A service quality and success model for the information service industry. European Journal of Operational Research, 156, 628–642.

Ledden L, Kalafatis SP, Samouel P. (2007) The relationship between personal values and perceived value of education. J. Bus. Res., 60(9), 965–974

Lee Ch, Ran B, Yang F, Loh WY. (2010) A Hybrid Tree Approach to Modeling Alternate Route Choice Behavior With Online Information. Journal of Intelligent Transportation Systems, 14(4), 209–219

Lehtinen U, Lehtinen J. (1982) Service Quality – A Study of Quality Dimensions, Service Management Institute, Helsingfors.

Lehtinen U, Lehtinen J. (1991) Two approaches to service quality dimensions. The Service Industries Journal, 11(3), 287-303.

Lewis RJ. (2000) An Introduction to Classification and Regression Tree (CART) Analysis. 2000 Annual Meeting of the Society for Academic Emergency Medicine in San Francisco, California.

Lin JH, Lee TR, Jen W. (2008) Assessing asymmetric response effect of behavioral intention to service quality in an integrated psychological decision making process model of intercity bus passengers: a case of Taiwan. Transportation, 35(1), 129–144.

Liou JH, Tzeng GH. (2007) A non-additive model for evaluating airline service quality. Journal of Air Transport Management, 13, 131–138.

Liou JJH, Tang C, Yeh W, Tsai C. (2011a) A decision rules approach for improvement of airport service quality. Expert Systems with Applications, 38(11), 13723-13730.

Liou JJH, Tsai C, Lin R, Tzeng G. (2011b) A modified VIKOR multiple-criteria decision method for improving domestic airlines service quality. Journal of Air Transport Management, 17(2), 57-61.

Liu J, Gao H. (2007) Study on railway transport service quality evaluation. 2007 International Conference on Wireless Communications, Networking and Mobile Computing, WiCOM 2007, 3308-3311.

López-Bonilla JM, López-Bonilla LM. (2008) Influence of the state-owned airlines on passenger satisfaction. Journal of Air Transport Management, 14(3), 143-145.

Martilla JA, James JC. (1977) Importance-performance analysis. Journal of Marketing, 14(1), 77–79.

Mathisen TA, Solvoll G. (2010) Service quality aspects in ferry passenger transport - examples from Norway. European Journal of Transport and Infrastructure Research, 10(2), 142-157.

Mattsson J. (1992) A service quality model based on ideal value standard. International Journal of Service Industry Management, 3(3), 18-33.

Minser J, Webb V. (2010) Quantifying the benefits: Application of customer loyalty modeling in public transportation context. Transportation Research Record, 2144, 111-120.

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.

Moons E,Wets G, Aerts M,Vanhoof K, Arentze TA, Timmermans HJP. (2001) The impact of irrelevant attributes on the performance of classifier systems in generating activity schedules. Electronic Proceedings of the 81st Annual Meeting of the Transportation Research Board Transportation Research Board,Washington, DC

Moons E, Wets G, Aerts M, Arentze TA, Timmermans HJP. (2005) The impact of simplification in a sequential rule-based model of activity-scheduling behavior. Environment and Planning A 37, 551–568.

Murray SJ, Walton D, Thomas JA. (2010) Attitudes towards public transport in New Zealand. Transportation, 37(6), 915-929.

Nathanail E. (2008) Measuring the quality of service for passengers on the hellenic railways. Transportation Research Part A, 42, 48–66

Nejati M, Nejati M, Shafaei A. (2009) Ranking airlines' service quality factors using a fuzzy approach: study of the Iranian society. International Journal of Quality & Reliability Management, 26(3), 247-260.

Nurul-Habib KM, Kattan L, Islaam T. (2011) Model of personal attitudes towards transit service quality. Journal of Advanced Transportation, 45, 271-285.

Opricovic S, Tzeng GH. (2004) Compromise solution by MCDM methods: A comparative analysis of VIKOR and TOPSIS. European Journal of Operational Research, 156(2), 445–455.

Ostrowski PL, O'Brien TV, Gordon GL. (1993) Service quality and customer loyalty in the commercial airline industry. Journal of Travel Research, 32, 16–24.

Oyewole P., (2001) Consumer's socio-demographic characteristics and satisfaction with services in the airline industry. Services Marketing Quarterly, 23(2), 61–80

Ozment J, Morash EA. (1998) Assessment of the relationship between productivity and performance quality in the U.S. domestic airline industry. Transportation Research Record, 1622, 22–30.

Pakdil F, Aydin Ö. (2007) Expectations and perceptions in airline services: An analysis using weighted SERVQUAL scores. Journal of Air Transport Management, 13, 229–237.

Pakgohar A, Tabrizi RS, Khalili M, Esmaeili A. (2011) The role of human factor in incidence and severity of road crashes based on the CART and LR regression: a data mining approach. Procedia Computer Science, 3, 764‒769

Paquette J, Bellavance F, Cordeau J, Laporte G. (2012) Measuring quality of service in dial-a-ride operations: The case of a canadian city. Transportation, 39(3), 539-564.

Parasuraman A, Berry LL, Zeithaml VA. (1991a) Refinement and reassessment of the SERVQUAL scale. Journal of Retailing, 67(4), 420-450.

Parasuraman A, Zeithaml VA, Berry LL. (1985) A conceptual model of service quality and its implications for future research. Journal of Marketing, 49, 41–50.

Parasuraman A, Zeithaml VA, Berry LL. (1988) SERVQUAL: a multi-item scale for measuring consumer perceptions of service quality. Journal of Retailing, 64(2), 12-40.

Parasuraman A, Zeithaml VA, Berry LL. (1991b) Understanding customer expectations of service. Sloan Management Review, 32(3), 39–48.

Parasuraman A, Zeithaml VA, Berry LL. (1994) Reassessment of expectations as a comparison standard in measuring service quality: implications for further research. Journal of Marketing, 58(1), 111-124.

Park J. (2007) Passenger perceptions of service quality: Korean and Australian case studies. Journal of Air Transport Management, 13, 238–242.

Park J, Robertson R, Wu CL. (2004) The effect of airline service quality on passengers' behavioural intentions: a Korean case study. Journal of Air Transport Management, 10, 435–439.

Park J, Robertson R, Wu CL. (2006) Modelling the impact of airline service quality and marketing variables on passengers' future behavioural intentions. Transportation Planning and Technology, 29(5), 359–381.

Pedersen T, Friman M, Kristensson P. (2011) Affective forecasting: Predicting and experiencing satisfaction with public transportation. J Appl Soc Psychol., 41(8), 1926-1946.

Peng Z, Luan X. (2011) Application research on traffic modal choice based on decision tree algorithm. Applied Mechanics and Materials, 97-98, 843-848

Philip G, Hazlett SA. (1997) The measurement of service quality: a new P-C-P attributes model. International Journal of Quality & Reliability Management, 14(3), 260-286.

Pullen WT. (1993) Definition and measurement of quality of service for local public transport management. Transport Reviews, 13(3), 247–264.

Quinlan JR. (1986) Induction of Decision Trees. Machine Learning, Vol. 1, 81–106.

Quinlan JR. (1993). C4.5 Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo (CA).

Quinlan, R. (1998). *C5.0: An Informal Tutorial*. RuleQuest. Available at http://www.rulequest.com/see5-unix.html

Ringle CM, Sarstedt M, Zimmermann L. (2011) Customer satisfaction with commercial airlines: The role of perceived safety and purpose of travel. Journal of Marketing Theory and Practice, 19(4), 459-472.

Ritchie JRB, Johnston EE, Jones VJ. (1980) Competition fares and fences perspective of the air traveler. Journal of Travel Research 18(3), 17–25

Saaty TL. (1994) How to make a decision: The analytic hierarchy process. Interfaces, 24, 19-43.

Saha GC, Theingi. (2009) Service quality, satisfaction, and behavioural intentions: A study of low-cost airline carriers in thailand. Managing Service Quality, 19(3), 350-372.

Sampson SE, Showalter MJ. (1999) The performance-importance response function: observations and implications. Service Industries Journal, 19(3), 1–25.

Sánchez M, Gázquez JC, Marín GM, Sánchez R. (2007) Effects of service quality dimensions on behavioural purchase intentions. A study in public-sector transport. Managing Service Quality, 17(2), 134-151

Sasser WE, Olsen RP, Wyckoff DD. (1978) Understanding service operations, Management of Service Operations, Ally & Bacon, Boston, MA.

Slack N. (1994) The importance-performance matrix as a determinant of improvement priority. International Journal of Operations & Production Management, 14(5), 59–75.

Smith AM. (1995) The consumer's evaluation of service quality: an examination of the ServQual methodology. Doctoral dissertation, University of Manchester, British Thesis Service, D189377.

Spreng RA, Mackoy RD. (1996) An empirical examination of a model of perceived service quality and satisfaction. Journal of retailing, 722, 201-214.

Stradling S, Anable J, Carreno M. (2007) Performance, importance and user disgruntlement: a six method for measuring satisfaction with travel modes. Transportation Research Part A, 41(1), 98–106.

Stuart KR, Mednick M, Bockman J. (2000) Structural equation model of customer satisfaction for the New York city subway system. Transportation Research Record, 1735, 133–137.

Sultan F, Simpson MC. (2000) International service variants: airline passenger expectations and perceptions of service quality. Journal of Services Marketing, 14(3), 188-216.

Surovitskikh S, Lubbe B. (2008) Positioning of selected middle eastern airlines in the south african business and leisure travel environment. Journal of Air Transport Management, 14(2), 75-81.

Teas RK. (1993) Expectations, performance evaluation, and consumers' perceptions of quality. Journal of Marketing, 57(4), 18-34.

Thill JC, Wheeler A. (2000). Tree induction of spatial choice behavior. Transportation Research Record 1719, 250–258.

Transport for London (2006) Network Performance of London Buses. Available at: http://www.tfl.gov.uk/buses/about/performance/service-report.asp).

Transportation Research Board (1999) A Handbook for Measuring Customer Satisfaction and Service Quality, TCRP Report 47.

Transportation Research Board (2004) Transit Capacity and Quality of Service Manual, second ed.

Tripp C, Drea JT. (2002) Selecting and promoting service encounter elements in passenger rail transport. The Journal of Services Marketing, 16(5), 432-442.

Tsai W, Hsu W, Chou W. (2011) A gap analysis model for improving airport service quality. Total Quality Management and Business Excellence, 22(10), 1025-1040.

Tsaur SH, Chang TY, Yen CH. (2002) The evaluation of airline service quality by fuzzy MCDM. Tourism Management, 23, 107–115.

Tyrinopoulos Y, Aifadopoulou G. (2008) A complete methodology for the quality control of passenger services in the public transport business. European Transport, 38, 1-16.

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

UNE-EN 13186 (2003) Transportation. Logistics and Services. Public Passenger Transport. Service Quality Definition, Targeting and Measurement. AENOR.

Van Dyke TP, Kappelman AL, Prybutok VR. (1997) Measuring information systems service quality: concerns on the use of SERVQUAL questionnaire. MIS Quarterly, 21(2), 195–208.

Vavra TG. (1997) Improving Your Measurement of Customer Satisfaction: A Guide to Creating, Conducting, Analysis, and Reporting Customer Satisfaction Measurement Programs (Milwaukee, WI: ASQ Quality Press).

Wang S, Feng C, Hsieh C. (2010) Stakeholder perspective on urban transport system service quality. Total Quality Management and Business Excellence, 21(11), 1103-1119.

Washington SP, Karlaftis MG, Mannering FL. (2003) Statistical and econometric methods for transportation data analysis. Florida: Chapman & Hall/CRC Press.

Washington S, Wolf J. (1997) Hierarchical tree-based regression: Theory and example applied to trip generation. Transportation Research Record 1581, Transportation Research Board,Washington, D.C., 82–88.

Washington S, Wolf J, Guensler R. (1997) Binary recursive partitioning method for modeling hot-stabilized emissions from motor vehicles. Transportation Research Record, 1597, 96-105

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

Wen CH, Lai SC, Yen W. (2008) Segmentation and positioning analysis for international air travel market: Taipei-to-Tokyo route. Transportation Research Record, 2052, 46-53.

Wen CH, Lai SC. (2010) Latent class models of international air carrier choice. Transportation Research Part E, 46, 211-221.

Wen CH, Lan LW, Cheng HL. (2005) Structural equation modeling to determine passenger loyalty toward intercity bus services. Transportation Research Record, 1927, 249–255.

Wets G, Vanhoof K, Arentze TA, Timmermans HJP. (2000). Identifying decision structures underlying activity patterns: an exploration of data mining algorithms. Transportation Research Record 1718, 1–9.

Wiley DE. (1973) The identification problem for structural equation models with unmeasured variables. In Structural Equation Models in the Social Science, A. S. Goldberger and O. D. Ducan, eds. New York: Seminar Press, 69–83.

Wipper LR. (1993) Performance measurement: producing results at the Oregon department of transportation. Transportation Research Record, 1395, 168–175.

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

Wolf J, Guensler R, Washington S, Bachman W. (1998) High-emitting vehicle characterization using regression tree analysis. Transportation Research Record, 1641, 58-65

Wong J, Chung P. (2007) Managing valuable Taiwanese airline passengers using knowledge discovery in database techniques. Journal of Air Transportation Management 13(6):362-370.

Xie C, Lu JY, Parkany E. (2003). Work travel mode choice modeling with data mining: decision trees and neural networks. Transportation Research Record 1854, 50–61.

Yamamoto T, Kitamura R, Fujii J, (2002) Driver's route choice behavior: analysis by data mining algorithms. Transportation Research Record 1807, 59–66.

Yang K, Hsieh T, Li H, Yang C. (2012) Assessing how service quality, airline image and customer value affect the intentions of passengers regarding low cost carriers. Journal of Air Transport Management, 20, 52-53.

Yeh CH, Deng H, Chang YH. (2000) Fuzzy multicriteria analysis for performance evaluation of bus companies. European Journal of Operational Research, 126, 459-473.

Yeh CH, Kuo YL. (2003) Evaluating passenger services of Asia-Pacific international airports. Transportation Research Part E, 39, 35–48.

Zadeh LA. (1965) Fuzzy sets. Information and Control, 8, 338-353.

Zeithaml VA, Berry LL, Parasuraman A. (1993) The nature and determinants of customer expectations of service. Journal of the Academy of Marketing Science, 12(1), 1–12.