Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 162 (2014) 459 – 466

Procedia
Social and Behavioral Sciences

XVIII Congreso Panamericano de Ingeniería de Tránsito, Transporte y Logística (PANAM 2014)

Cluster analysis for diminishing heterogeneous opinions of service quality public transport passengers

Rocio de Oña*, Griselda López, Fco Javier Díez de los Rios, Juan de Oña

*TRYSE Research Group, Department of Civil Engineering, University of Granada,
ETSI Caminos, Canales y Puerto, c/ Severo Ochoa, s/n, 18071 Granada (Spain)*

Abstract

One of the principal measures that public transport administrations are following for reaching a sustainable transportation in the cities consists on attract a higher number of citizens towards the use of public transport modes, by offering high quality services. Collecting users opinions is the best way of detecting where the service is failing and which aspects are been provided successfully. The main problem that has to be faced for analyzing service quality is the subjective nature of its measurement, offering heterogeneous assessments among passengers about the service. Stratifying the sample of users on segments of passengers which have more uniform opinions about the service can help to reduce this heterogeneity. This stratification usually is conducted based on the social and demographic characteristics of the passengers. However, there are more advance techniques that permits to identify more homogeneous groups of users. One of these techniques is the Cluster Analysis, which is a data mining technique that can be used for segmenting the sample of passengers on groups that share some common characteristics, and that have more homogeneous perceptions about the service. This technique has been applied in other fields of transport engineering but it has never been applied for searching homogeneous groups of users with regards to service quality evaluation in a public transport service. For this reason, the aim of this work is to find groups of passengers that perceive the quality of the service in a more homogeneous way, and to apply to this clusters a suitable statistic technique that permit us to discover which are the variables that more influence the passengers' overall evaluation about the service. The comparison among the results of each cluster will show considerable differences among them and also with the results obtained using the global sample.

Keywords: Service Quality; Public transportation, Cluster analysis, Pearson correlation

* Corresponding author. Tel.: +3-495-824-9455; fax:+3-495-824-6138.
E-mail address: rociodona@ugr.es

doi:10.1016/j.sbspro.2014.12.227

1. Introduction

As the use of public transport modes represents a solution to the cities' environmental and social problems generated by the citizens'mobility (pollution, traffic congestion, noise, etc.), the definition and measurement of service quality in public transportation becomes essential as a tool for increasing the attractiveness of this more sustainable alternative. For individuals, car travel is generally perceived as more comfortable, flexible and faster for supporting busy lifestyles (Jakobsson Bergstad et al., 2011).That is why public transport services have to prove themselves their competition with private vehicle and become the support for the movement of passengers, instead of act only as a possible alternative.

Thus, Transport Authorities are focused on impose strong incentives for guaranteeing effective and high quality public transport services. Due to passengers are who suffer poor services or high levels of performance, it seems logical that their opinions will be crucial for analyzing the service. In fact, the Transit Capacity and Quality of Service Manual (Transportation Research Board, 2004) considers the customer point of view the fundamental perspective for service assessment.

Therefore, Customer Satisfaction Surveys (CSSs) are usually used for collecting and process passengers' opinions in order to design and formulate adequate interventions and strategies (de Oña et al 2012; 2013a; 2014b). The main problem up until now of the analysis is the fuzzy and heterogeneous assessment among passengers. One solution to help to reduce this heterogeneity can be to stratify the sample of users on segments of passengers more homogeneous. Dell'Olio et al. (2010) developed this stratification in order to propose specific models for analyzing service quality. Likewise, de Oña et al. (2014a) stratified the sample of a railway service operating in the north of Italy for the same purpose. The base used for conducting this stratification was the social and demographic characteristics of the passengers (i.e. models for men or women, for the younger, according to income level, etc.), or their travel habits profiles (i.e. type of the day of the journey, time of the day, frequency of use, etc). However, using this methodology, the heterogeneity among users could still be presented.

There are more advance techniques, such as Cluster Analysis (CA), which permit to reduce this heterogeneity, by segmenting the sample of passengers on groups that share some common characteristics, and that have more homogeneous perceptions about the service. This technique has been applied in other fields of transport engineering with satisfactory results (Karlaftis & Tarko, 1998; Ma & Kockelman, 2006; Pardillo-Mayora, 2010; Depaire et al., 2008; de Oña et al., 2013b). The main purpose of this study will be to apply a cluster analysis technique for stratifying the sample of users of a public transport service in the city of Granada (Spain), in order to analyze passengers' opinions under more homogeneous conditions. Data from four Customer Satisfaction Surveys developed from 2008 to 2011 in the metropolitan public bus service of Granada are used, and the most important variables affecting service quality are determined using a Pearson Correlation. Furthermore, the differences found among the key factors influencing the service quality evaluation of the identified groups of passengers are displayed and discussed.

The paper is structured as follows: section 2 shows the methodology used for stratifying the sample and for evaluating service quality. Section 3 describes data used for the analysis. Next, the results obtained with the cluster analysis and the Pearson correlation will be explained, and finally the conclusions are reported.

2. Techniques and procedures

2.1. Analysis cluster

Analysis cluster may be defined as statistical classification technique in which cases, data, or objects (events, people, things, etc.) are sub-divided into groups (clusters) such that the items in a cluster are very similar to one another and very different from the items in other clusters.

Latent Class Clustering (LCC) is a particular method to build cluster, which permits use frequencies, categorical and metric variables. In addition, LCC does not need a prior standardization of the data that could have a bearing on the results. (Magidson & Vermunt, 2002; Vermunt & Magidson, 2005).

A model of LCC can be expressed as follow: given a data sample of N cases, measured with a set of observed variables, Y1,…,Yj , which are considered indicators of a latent variable X; and where these variables form a Latent

Class Model (LCM) with T classes. If each observed value contains a specific number of categories: Yi contains $\mathrm{I_i}$ categories, with i=1…j; then the manifest variables make a multiple contingency table with $\prod_{i=1}^{j} I_i$ response patterns. If π denotes probability, $\pi(X_t)$ represents the probability that a randomly selected case belongs to the latent t class, with t=1, 2,…, T.

The expression of LCMs is given by (1):

$$\pi_{Y_i} = \sum_{t=1}^{T} \pi_{X_t}\pi_{Y_i|X_t}, \tag{1}$$

with Y_i response-pattern vector of case i; $\pi(X_t)$ is the prior probability of membership in cluster t; $\pi_{Y_i|X_t}$is the conditional probability that a randomly selected case has a response pattern Y_i= (y1,…,yj), given its membership in the t class of latent variable X. The assumption of local independence needs to be verified, and therefore Eq. (1) is re-written:

$$\pi_{Y_i} = \sum_{t=1}^{T} \pi_{X_t} \prod_{i=1}^{j} \pi_{Y_{ij}|X(t)}, \text{ with } \sum_{i=1}^{j} \pi_{Y_{ij}|X(t)} = 1, \text{ and } \sum_{t=1}^{T} \pi_{X_t} = 1, \tag{2}$$

The estimation of the model is based on the nature of the manifest variables, since it is assumed that the conditional probabilities may follow different formal functions (Vermunt & Magidson, 2005). The method of maximum likelihood is used for estimating the model's parameters. Once the model has been estimated, the cases are classified into different classes by using the Bayes rule to calculate the a posteriori probability that each n subject comes from the t class (are the model's estimated values):

$$\pi_{X_t|Y_i} = \frac{\hat{\pi}_{X_t}\hat{\pi}_{Y_i|X_t}}{\hat{\pi}_{Y_i}}, \tag{3}$$

In practice, the set of probabilities is calculated for each response pattern and the case is assigned to the latent case in which the probability is the highest. Thus, a specific passenger may belong to different latent cases with a specific percentage of membership (with 100% being the sum total of membership probabilities).

A priori, the number of cluster is unknown, therefore the aim is to find the model that can explain or adapt the best to the data being used. LCC deals with model selection (number of clusters) by trying multiple models and computing various information criteria such as the Bayesian Information Criteria (BIC) (Raftery, 1986), Akaike Information Criterion (AIC) (Akaike, 1987), and Consistent Akaike Information Criterion (CAIC) (Fraley & Raftery, 1998). The appropriate number of clusters is the one that minimizes the score of these criteria, because the model is more parsimonious and adapts better to the study data (de Oña et al., 2013b).

2.2. Pearson Correlation

The Pearson correlation developed by Karl Pearson (Person, 1985) is a tool used for estimating the linear relationship between two quantitative random variables. This methodology is independent of the scale of measure of the variables. Then, when two random variables are studied (x and y) on a statistic population sample; the coefficient of the Pearson correlation is denominated as $\rho_{x,y}$, and it is calculated with the following expression (4):

$$\rho_{x,y} = \frac{cov(X,Y)}{\sigma_x\sigma_y} = E\frac{[(X-\mu_x)(Y-\mu_y)]}{\sigma_x\sigma_y}, \tag{4}$$

Where, $\sigma_{x,y}$ is the (X, Y) covariance, σ_x is the X variable typical deviation, and σ_x is the Y variable typical deviation.

The correlation index value can move between [-1, 1]. When the value is positive, it exist a positive correlationship between the variables, then if one of the variables increases, the other one increases too. On the opposite, a negative value will indicate that if a variable increases its magnitude the other one will decrease its value. Coefficients -1 and 1 represent a perfect negative and positive correlation respectively, versus a coefficient $\rho_{x,y}$= 0 that indicatesno lineal relationship between the variables.

3. Data of study

This study involves 3,664 interviews collected in four consecutive CSSs developed from 2008 to 2011 (around 1,000 face-to-face surveys are conducted annually) in the bus metropolitan public transport service of Granada (Spain). The CSSs are divided into four main sections:

- The first section is about general information of the trip, with regards to time of the interview, the bus stop, the line, origin, destination, etc.
- Second section refers to the passengers' travel habits: the reason for travelling, the frequency of use, the type of ticket, the availability of private vehicle, the complementary modes used from the origin to the bus stop and from the bus stop to the destination, etc.
- The socioeconomic characteristics of the passengers are required in the third section (sex and age).
- And finally, the last section of the survey is specifically about passengers' perceptions about the service characteristics. First, the interviewers asked the passengers about their perception of performance with regards to 12 Service Quality (SQ) factors, on a cardinal scale from 0 to 10. Second, they asked the passengers to identify the three most important SQ factors for each of the 12 factors. And finally, they asked about the overall SQ perception based on a cardinal scale from 1 to 5. The variables used to measure the perception of the SQ attributes included information, punctuality, safety on board, driver courtesy, bus interior cleanliness, bus space, bus temperature, accessibility to/from the bus, fare, speed, frequency of service and stops proximity to/from origin/destination.

Table 1. Characteristics of the sample.

Variable	Categories	Statistics
Sex	1.Male	(32%) (n. obs.)
	2.Female	(68%)
Age	1.{18-30}	(49%)
	2.{31-60}	(40%)
	3.{>60}	(11%)
Travel reason	1. Job	(28%)
	2. Studies	(25%)
	3. Others	(47%)
Use	1. Frequent	(77%)
	2. Occasionally	(23%)
Type of ticket	1. Standard ticket	(23%)
	2. Consortium Pass	(67%)
	3. Senior Citizen Pass	(7%)
	4. Other	(3%)
Possession of private vehicle	1. Yes	(47%)
	2. No	(53%)
Trips from origin to bus stop	1. On foot	(77%)
	2. Vehicle	(23%)
Trips from bus stop to destination	1. On foot	(95%)
	2. Vehicle	(5%)

The characteristics of the collected sample are represented in Table 1. It is made up more of females than males. A little part of the respondents were older than 60, while the rest is handed out balanced in the other two groups. Likewise, it shows that the main reasons for travelling are job and studies, however, lot of other reasons are the purpose of their trips. Most of the passengers travel almost every day (more than four times a week), followed by passengers travelling frequently (from 1 to 3 times a week).The consortium pass is the type of ticket most used on the opposite to the standard ticket, the senior citizen pass and other type of ticket. The sample of users is equally distributed among those who had a private vehicle available for making the trip with those who did not have it available. Most part of the respondents access to the bus service on foot (77% of the passengers), while others use other type of mode (urban bus, metropolitan bus, private vehicle, motorbike, bicycle, taxi or others). Likewise, almost all the respondent access to their destination from the bus stop on foot.

4. Results

4.1. Stratification by Cluster analysis

The first step was to create the cluster, table 1offers a description of the variables used in the analysis. To select the number of clusters in the final model, different models in which the number of the cluster was varying (between 1 to 10) were tested. In order to select the number of cluster in the final model, the parameters BIC, AIC and CAIC were calculated. Then, it was obtained that increasing the number of clusters until four, the values of BIC, AIC and CAIC reduce, then the best model is the one with 4 clusters. In addition, the value of the entropy for model 4 is 0.766, which indicates a good separation between clusters (McLachlan & Peel, 2000).

In the final model, formed by 4 clusters, each cluster was characterized by the proportion of each variable in each cluster. Following the work of (Depaire et al., 2008 & de Oña et al., 2013b), the clusters were analyzed and named based on their variable distributions. Then, the most important categories within each cluster for each variable were identified (using for that the highest conditional probability obtained for a determined category of a variable given its membership to a specific cluster). Therefore, using this criterion the clusters were named.

Some variables cannot be used in the characterization of the clusters because the highest value of probability was obtained for the same category of the specific variable in all of the clusters built (for example “trips from origin to bus stop” and “trips from bus stop to destination”, in which the category selected is “going on foot”). Then, this variable does not permit a differentiation between the clusters.

The variables used to name the cluster and their probabilities in each one of the 4 clusters are represented in Table 2.The results are the following:

- Cluster 1: This cluster concentrates most of the data (39% of the data). It is mainly formed by women (64% of the data). Likewise, most of them are young people (95% of the data) and they use the consortium pass (86% of the data). This peoples use the transport very frequent (99% of the data) and their reason for travelling is studies (68% of the data). Finally, stand out that they do not have available a private vehicle (61% of the data). Sum up, they are woman who use the public transport frequently because studies reasons, and they do not have a private vehicle available.
- Cluster 2: The size of this cluster is 28% of the data. It is mainly formed by women (80% of the data)with a bigger percent than cluster 1. Likewise, most of them have medium age (59% of the data) They are frequent users in the 99% using the consortium pass (90%). About the travel reason it is occupation with a 62% of probability.
- Cluster 3: The size of this cluster is 23% of the data. It is mainly formed by women (64% of the data), same to cluster 1 and 2. Most of them have medium age (59% of the data) and they use the standard ticket (65% of the data). The users are sporadic with a probability of the 87%, and the travel reason is other (88%).
- Cluster 4: This is the cluster of minor size, with 9% of the data. It is formed by women and men (43% are men and 57% are women), being old users (> 60 years old) in the 99% of the cases. They do not have a private vehicle (77% of probability). About the ticket used is the senior citizen pass (78%), and the travel reason is other (in the 99% of the cases).

Table 2. Variables, categories and probabilities of membership in the cluster.

Variables	Category	Cluster1	Cluster2	Cluster3	Cluster4
Possession of private vehicle	No	**61%**	47%	43%	**77%**
	Yes	39%	**53%**	**57%**	23%
Travel reason	Job	16%	**62%**	11%	1%
	Studies	**68%**	0%	2%	0%
	Others	16%	38%	**87%**	**99%**
Use	Frequent	**99%**	**99%**	32%	**50%**
	Occasionally	1%	1%	**68%**	**50%**
Ticket	Standard	11%	9%	**65%**	7%
	Conscard	**86%**	**90%**	28%	13%
	Fasscard	0%	0%	0%	**78%**
Age	Young	**95%**	20%	37%	0%
	Middle	5%	**78%**	**59%**	1%
	Old	0%	2%	4%	**99%**
Sex	Men	36%	20%	36%	43%
	Women	**64%**	**80%**	**64%**	**57%**
Total		1435	1033	859	261

4.2. Importance of the variables by Pearson correlation

A Pearson correlation has been applied for each cluster in order to derive the importance of the service quality attributes in each of the groups identified. For this aim, it is checked the lineal relationship between each attribute with the "overall SQ evaluation". Table 3 displays the obtained Pearson coefficients for each Cluster.

Cluster 1: In this cluster, the punctuality and frequency are the attributes more related with the overall SQ evaluation, so they could be defined as the most important variables for this group of passengers. The results are according with the group analyzed: young students that are women and do not have private car. So, for them, the public transport service is necessary, and due to their not flexible timetable because of the lessons, the punctuality becomes a high priority service characteristic.

Cluster 2: In this cluster, the frequency and the information are the most important variables. This cluster group is formed by women who have medium age and they use frequently the public transport service for reaching the job. This group needs a high service frequency because, on the contrary to the students, their timetable is more flexible so it is preferable to have a more continuous service than its punctuality. Information is also placed on a high position of the ranking due to the high sensitivity of frequent users to the disturbances of the regular performance of the service, being who more suffer these changes because of their frequent use.

Cluster 3: In this cluster, the frequency, the information and the speed are the most important variables. The main different with the cluster before is that in this case women have the possibility of driving a private vehicle for making the trip, and their frequency of use is occasional. That is why the speed is considered an essential factor of the service for carrying out their modal choice, due to the trip duration using the public transport service should be competitive with the one took by the private vehicle. If not, maybe the public transport service does not represent a modal alternative for this group of passengers.

Cluster 4: In this cluster, the variables highlighted are the information, the frequency and the accessibility. This cluster group is formed with old people who use the public transport for other reasons, not job or studies. These results have sense as usually for elderly people the used of new technologies represent a challenge, so they look for a clear and intuitive information about the performance of the service. Moreover, the accessibility to and from the vehicle is decisive for them because sometimes they have a reduced mobility that becomes worse over the years.

Table 3. Classification of variables according to Pearson Correlation.

Cluster 1		Cluster 2		Cluster 3		Cluster 4	
Attribute	P.C.C.	Attribute	P.C.C.	Attribute	P.C.C.	Attribute	P.C.C.
Punctuality	0.345**	Frequency	0.465**	Frequency	0.383**	Information	0.421**
Frequency	0.331**	Information	0.432**	Information	0.336**	Frequency	0.348**
Space	0.328**	Punctuality	0.392**	Speed	0.323**	Accessibility	0.327**
Information	0.321**	Temperature	0.383**	Punctuality	0.314**	Punctuality	0.276**
Safety	0.305**	Fare	0.368**	Safety	0.298**	Speed	0.255**
Speed	0.296**	Space	0.367**	Courtesy	0.283**	Safety	0.251**
Temperature	0.291**	Speed	0.360**	Temperature	0.273**	Temperature	0.240**
Courtesy	0.266**	Safety	0.350**	Proximity	0.264**	Proximity	0.239**
Accessibility	0.258**	Accessibility	0.345**	Fare	0.221**	Cleanliness	0.229**
Fare	0.253**	Cleanliness	0.336**	Space	0.221**	Space	0.205**
Proximity	0.239**	Courtesy	0.301**	Cleanliness	0.216**	Courtesy	0.171**
Cleanliness	0.236**	Proximity	0.292**	Accessibility	0.206**	Fare	0.169**

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

5. Conclusions

In this study an analysis of service quality in a metropolitan public bus service of Granada was conducted by using Cluster Analysis and a Pearson Correlation. Data from various customer satisfaction surveys carried out over the period 2008-2011 were analyzed. Cluster Analysis was used for identifying different profiles of users that have more homogeneous opinions about the service. This advance segmentation technique is able to consider at the same time various socioeconomics characteristics of the users and their travel habits for finding the groups or clusters. Subsequently, the key factors influencing service quality evaluation were identified by using a Pearson Correlation. A ranking was built with this correlation, and it showed significant differences across the groups of passengers, being the order of this ranking different according to the group under consideration.

Then, the outcomes find out interesting insights. Cluster analysis determined four groups of passengers, representing diverse profiles. Cluster 1 defines a passenger that is women, young, that travel for studies reasons, that their trips are frequent, using a consortium pass, and they do not have available a private vehicle. For this sort of passenger, the most important variable was the Punctuality, maybe because they have to arrive on time to the lessons and exams. In the case of Cluster 2, represented by middle age women, travelling frequently for occupation reasons and using the consortium pass, the most important variables were the Frequency and the Information. The timetable of working people usually is more flexible than for students, so a higher quality frequency would be preferred to the Punctuality. On the rest of Clusters (3 and 4) also the variables that played an important role in passengers overall evaluation varied.

The findings brought up in this research prove that passengers' opinions are very heterogeneous among them, and that personalized analysis is a good solution for diminishing this heterogeneity, in order to recognize the influence that the variables have on different profiles of passengers. Some information and some details about service quality evaluation could be masked if a global treatment of the data is developed.

These issues are of interest for transport planners, who, in order to formulate successful incentives for promoting the public transport service, they should decide which users they want to engage, for attending their preferences and needs and apply a personalized strategy.

Acknowledgements

This study is sponsored by the Consejería de Innovación, Ciencia y Economía of the Junta de Andalucía (Spain) through the Excellence Research Project denominated "Q-METROBUS-Quality of service indicator for METROpolitan public BUS transport services". The authors also acknowledge the Granada Consorcio de Transportes for making the data set available for this study. Likewise, Griselda López wishes to express her acknowledgement to the regional ministry of Economy, Innovation and Science of the regional government of Andalusia (Spain) for their scholarship to train teachers and researchers in Deficit Areas

References

Akaike, H., (1987). Factor analysis and AIC. *Psychome,* 52, 317–332.

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

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

de Oña J, de Oña R, Eboli L, Mazzulla G. (2013a). Perceived service quality in bus transit service: A structural equation approach. Transport Policy 29:219-226

de Oña, J., de Oña R, Eboli L, Mazzulla G. (2014a). Heterogeneity in perceptions of service quality among groups of railway passengers. International Journal of Sustainable Transportation (DOI:10.1080/15568318.2013.849318)

de Oña, J., López, G., Mujalli, R. O., Calvo, F. J. (2013b). Analysis of traffic accidents on rural highways using Latent Class Clustering and Bayesian Networks. *Accident Analysis and Prevention*, 51, 1–10.

de Oña R, Eboli L, Mazzulla G. (2014b). Key factors affecting rail service quality. a decision tree approach. Transport (DOI:10.3846/16484142.2014.898216)

Depaire, B., Wets, G., Vanghoof, K. (2008). Traffic accident segmentation by means of latent class clustering. *Accident Analysis and Prevention,* 40 (4), 1257–1266.

Fraley, C., Raftery, A.E., (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal,* 41, 578–588.

Jakobsson Bergstad, C., Gamble, A., Hagman, O., Polk, M., & Garling, T. (2011). Affective-symbolic and instrumental-independence psychological motives mediating effects of socio-demographic variables on daily car use. *Journal of Transport Geography,* 19 (1), 33–38.

Karlaftis, M., Tarko, A., (1998). Heterogeneity considerations in accident modeling. *Accident Analysis and Prevention,* 30 (4), 425–433.

Ma, J., Kockelman, K., (2006). Crash frequency and severity modeling using clustered data from Washington state. In: IEEE Intelligent Transportation Systems Conference, Toronto, Canada.

Magidson, J., & Vermunt, J.K., (2002). Latent class models for clustering: A comparison with K-means. *Canadian Journal of Marketing Research,* 20, 37-44.

Mclachlan, G.J., Peel, D., (2000). Finite Mixture Models. Wiley, New York.

Pardillo-Mayora, J.M., Domínguez-Lira, C.A., Jurado-Piña, R., (2010). Empirical calibration of a roadside hazardousness index for Spanish two-lane rural roads. *Accident Analysis and Prevention,* 42, 2018–2023.

Pearson, K. (1895). Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58, 240 – 242.

Raftery, A.E., (1986). A note on Bayes factors for log-linear contingency table models with vague prior information. *Journal of the Royal Statistical Society*, Series B 48, 249–250.

Transportation Research Board, (2004). Transit Capacity and Quality of Service Manual, second ed.

Vermunt, J. K., & Magidson, J. (2005). Latent GOLD 4.0 User Manual. Belmont Ma.: Statistical Innovations Inc.