SELECTING KEY QUALITY INDICATORS IN PUBLIC TRANSPORT SYSTEMS USING A ROBUST METHOD

BENEDETTO BARABINO[1], CLAUDIO CONVERSANO[2],
NICOLA ALDO CABRAS[2] & MASSIMO FANTOLA[1]
[1]Department of Civil Engineering, Environment and Architecture, University of Cagliari, Italy
[2]Department of Business and Economics Sciences, University of Cagliari, Italy

ABSTRACT

Recent interests in transit services have captured attention of experts on the monitoring of public transport quality. Previous research has focused on the development of relevant models and methods for monitoring transit service quality and identified where and when different levels of service quality occur. However, little attention has been given to objectively determining a pool of key quality indicators (KQIs) for monitoring purposes. In this paper, we address this gap by proposing a robust methodology that identifies a lengthy list of potential KQIs and defines their properties, then enlists the judgement of both researchers and practitioners regarding each KQI, evaluates the initial list, and recommends the most promising KQI set. We demonstrate the effectiveness of this methodology using an international survey and Monte Carlo simulation methods. The outcomes will prove useful in helping practitioners to monitor transit service quality according to a set of recognised KQIs.
Keywords: key quality indicators, transit service quality monitoring, quality robust methodology.

1 INTRODUCTION

The term "quality of experience" (QoE) refers to the degree of satisfaction or disappointment a user experiences when using an application or service. QoE results from the fulfilment of a user's expectations with respect to the utility of and/or enjoyment derived from the application or service in light of the user's personality and current state [1]. In the transit service context, QoE is a broad and complex concept that can be defined as the capability of transit operators to provide a service based on the desires of current and potential users. It can be described by several indicators that represent various features of transit services. The interest in indicators is due to their ability to provide informative signals regarding areas of excellence and service priority in transit services. Moreover, indicators can represent a strategic aspect of the desired quality and serve as data input for the next step in the service quality monitoring process *i.e.,* measurement and management. Therefore, the selection of key quality indicators (KQIs) is crucial in order to evaluate the performance on quality of transit services, and presents four key challenges.

First, there are many potential service quality indicators in the transit service context, so selecting a compact pool of key indicators can be challenging [2].

Second, previous research on transit service quality has focused primarily on the development of models and methods for "managing" indicators rather than on the determination of what to measure [3]. More precisely, previous researchers have rarely used objective methods to develop indicator selection mechanisms. That is, sets of suitable quality indicators used to perform analysis have previously been selected as follows:

- Based on the literature and/or specific knowledge of various economic and industrial fields [4]–[9].
- Surveys of operators, users and focus groups [10]–[12].
- Specific organisational objectives [13], [14].
- Ad hoc theoretical methodologies [15], [16].

Third, since indicators are merely abstractions of the system and will never perfectly represent it, it is challenging to select the most appropriate indicators that will describe the system most effectively.

Fourth, a careful indicator selection process can improve their acceptability and credibility among experts. This enhanced acceptability is referred to as method robustness.

In this paper, we propose a robust methodology for identifying and selecting KQIs using both data collected through international surveys and Monte Carlo simulation methods. This methodology builds on the first block of TRANSQUAL, which recommends the use of a robust methodology for KQIs selection [15]. This methodology differs from ELASTIC, which is a framework for the identification and selection of sustainable transport indicators developed by UK experts only [17]. Finally, it facilitates and advances primary research by providing a compact pool of KQIs rather than the five KQIs reported in [10].

The remainder of this paper is organised as follows: In Section 2, we propose a method for identifying and selecting the most suitable KQIs for describing transit service quality. In Section 3, we present and discuss the application of this method and its results. Finally, in Section 4, we draw our conclusions and suggest future research directions.

2 METHODOLOGY

In this section, we propose a robust methodology for the identification and selection of a pool of KQIs. The methodology comprises six steps, as shown in the schematic of Fig. 1. A concise description of these steps is as follows.

2.1 Identification of a long list of potential quality indicators

First, a lengthy list of KQIs is identified. Several sources may be considered. In this method, relevant literature is identified with respect to both research and practice, such that a more comprehensive initial source of KQIs is obtained.

2.2 Definition of components and attributes of KQIs

In the second step, the components and attributes of the KQIs are established to select a pool of KQIs that are determined to be the most appropriate for describing transit service quality. The overall goal of this step is to provide effective direction for the evaluation process.

To achieve this goal, we consider two manageable components:

- The methodological features of the indicator.
- The relevance of the indicator to the concept of service quality.

These components, while more specific than the overall goal, may still be too broad to guide the selection process. Therefore, we must further decompose these components into more manageable attributes. According to [15], the methodological attributes are as follows:

- Measurability for considering the theoretical aspects of each indicator.
- Ease of availability for collecting data at a reasonable cost.
- Speed of availability for making regular assessment updates.
- Interpretability to avoid ambiguous or confusing outcomes.

The same reference [15] documents that the relevance of indicators to the concept of service quality includes four main attributes:
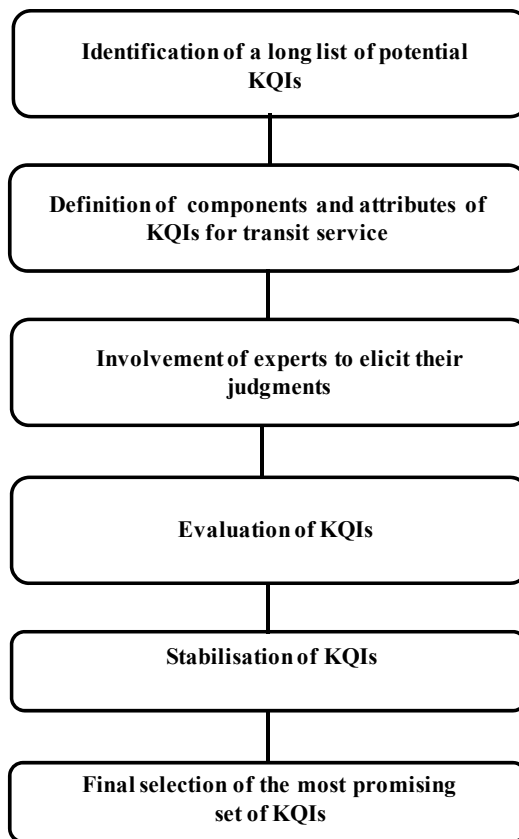
Figure 1:   Proposed robust methodology for identifying and selecting KQIs for monitoring transit service quality.

- Integration of users' and public transport company (PTC)'s views to re-orient the organisation as one that is customer-centred.
- User-orientation to avoid PTC-oriented parameters.
- Subjective and objective measurement to evaluate the impact on the satisfaction on the performance.
- Statement of the amount of passengers to prioritise actions.

   Methodological attributes are general aspects of the application of the method, whereas quality attributes are more specific in that they account for how well they reflect the impact on users of crucial aspects of service quality. In addition, the level of importance assigned to each component and attribute can differ to reflect the viewpoints of experts.

2.3 Involvement of experts to elicit their judgements

In the selection of KQIs, the third step is key in that it requires the involvement of academic researchers who can guide the concept of quality and practitioners who work every day to ensure service quality. Hereinafter, we refer to researchers and practitioners as experts.

Several approaches can be used for the involvement of experts. Experts can participate in focus groups (FGs) or surveys, depending on the amount of data required. FGs provide qualitative data collected in a workshop-style methodology using an established number of selected individuals. The main objective of FGs is to obtain information based on the opinions of the participants regarding the issue being investigated [10], [18]. Surveys provide quantitative data collected from a large pool of users, and there are many survey types (e.g., intercept, at home, e-mail, web and so on). Although FGs are an interesting tool for this purpose, their use is confined to a very restricted number of experts. Therefore, we propose the use of international surveys involving a range of experts. Regarding the type of survey used, we suggest the web-survey as it: 1) is able to elicit information at a low cost; 2) can rapidly reach experts; 3) can be compiled directly on-line, thus removing the need to print the questionnaire; and 4) results in data that are ready for processing.

In this methodology, experts need to be involved twice when gathering data input for the evaluation of each indicator. First, the experts help in prioritising the levels of importance of the components and attributes, with the recognition that these levels of importance will likely vary among experts. Thus, weights of importance are attached to these components and attributes (i.e., items) to reflect the viewpoints of experts. Although weights can be directly attached by questioning experts about their preferences on single items, this approach is imperfect as humans have difficulty transforming relevant information about many items into relative weights [19]. Hence, in this study, we used the analytical hierarchy procedure (AHP) [20], [21]: rather than assigning weights directly, weights are established based on a pairwise comparison of items. More precisely, the AHP elicits subjective comparisons by experts on pairs of items and then aggregates these outcomes into objective weights. The AHP is performed twice: once for the components and again for the attributes.

Second, the experts make their evaluations by giving marks for each indicator against each attribute according to predefined scales (e.g., numerical, qualitative and so on).

2.4 Evaluation of KQIs

In the fourth step, we compute the *Weighted Quality Indicator Performance Score* (*WQIPS*) for each KQI, by aggregating their previously derived weights and outcome marks.

More precisely, let:

- I be the set of indicators;
- J be the set of experts interviewed;
- H be the set of the methodological attributes;
- K be the set of the quality attributes;
- $\overline{m}$ be the average weight of the methodological component;
- $\overline{q}$ be the average weight of the relevance of quality component;
- $\overline{w_h}$ be the average weight of attribute $h \in H$;
- $\overline{z_k}$ be the average weight of attribute $k \in K$;
- $\overline{V_{\iota h}}$ be the average mark of indicator $i \in I$ for attribute $h \in H$;
- $\overline{V_{\iota k}}$ be the average mark of indicator $i \in I$ for attribute $k \in K$.

For each parameter *i*, the *WQIPS$_i$* is computed as follows:

$$WQIPS_i = \overline{m}\left(\sum_{h=1}^{|H|} \overline{w_h}\,\overline{V_{\iota h}}\right) + \overline{q}\left(\sum_{k=1}^{|K|} \overline{z_k}\,\overline{V_{\iota k}}\right) \forall\, i = 1\dots|I|. \qquad (1)$$

2.5 Stabilisation of KQI

The fifth step is necessary to ensure the stability of the computed $WQIPS_i$s, particularly when the number of answers provided by experts in the survey is not very large, as often happens in these kinds of studies. Stabilisation (or "robustification") of the estimated $WQIPS_i$ is performed using Monte Carlo simulation methods, which allow the analyst to quantify the effect of random variation, lack of knowledge or error on the sensitivity, performance or reliability of the system being modelled. The deterministic model of expert judgements is iteratively evaluated using a set of random numbers as inputs. In this way, the observed (deterministic) model is transformed into a stochastic model. Typically, the theoretical model used in the derivation of $WQIPS_i$ is simulated to produce artificial data that must be matched with observed data (i.e., judgements made by experts). Matching is performed by choosing a suitable probability distribution that best represents the current state of knowledge. Specifically, to derive a robust set of $WQIPS_i$s, we apply the Monte Carlo method twice to retrace the process designed in the theoretical model, which leads to the computation of the $WQIPS_i$ by eqn (1). Thus, two Monte Carlo experiments are performed, namely:

- Experiment 1: We simulate the distribution of the scores obtained for the entire set of $|H|+|K|$ items characterising the AHP many times B by generating multinomial distributions with $|H|+|K|$ classes. The choice of multinomial distribution is motivated by the consideration that, in most cases, $|H|+|K| > 2$. As such, a multi-class reference distribution is required. Of course, for the case $|H|+|K| = 2$, the shift to a binomial distribution is straightforward. The scores obtained from the pairwise comparisons of items by the experts are used as a priori probabilities of the $|H|+|K|$ classes in the data generation process (DGP). The DGP is based on a number of trials (sample size) corresponding to the number of participants involved in the survey and is simulated B times. The final estimated weights for each component and attribute are the averages of their mean values obtained in each trial.
- Experiment 2: We repeat the process used in Experiment 1 for the set of indicators related to the $|H|+|K|$ attributes. Usually, I is very large and the total number of items considered in the simulation is $S = (|H|+|K|) \times |I|$. For each indicator $i \in I$, the distribution of scores obtained by the experts is simulated B times using a binomial distribution X defined in the [0,10] interval, such that $E(X) = \overline{V_{th}}/|J|$ or $E(X) = \overline{V_{tk}}/|J|$ if indicator $i \in I$ concerns an attribute of a methodological or quality component, respectively. More precisely, the mean score obtained for an indicator $i \in I$ is rescaled on the [0,1] interval. Moreover, this is used as a parameter of the binomial distribution simulated for the number of trials that correspond to the number of items characterising a specific indicator. Again, the average score obtained for each indicator is derived by averaging the scores obtained in the various simulation trials.

2.6 Final selection of KQI

In the last step of the proposed methodology, indicators are ranked in decreasing order based on their estimated $WQIPS_i$s. The best indicator $i \in I$ is the one with the highest estimated $WQIPS_i$. More precisely, a double ordering of indicators is defined. In the first ordering, the $WQIPS_i$s are recomputed by replacing empirical weights with those obtained from Experiment 1. Next, they are sorted into decreasing order. In the second ordering, the $WQIPS_i$s are recomputed by replacing their empirical marks with those obtained from

Experiment 2. The final rank assigned to each individual *WQIPS*_i is the average rank obtained from the double ordering. The output of this process is the "robust" selection of a pool of top *n* indicators that are most representative of transit service quality.

3   EXPERIMENTAL RESULTS AND DISCUSSION

We implemented this methodology as shown in Fig. 1.

3.1   Preliminary setup

The European Committee for Standardisation (CEN)/TC 320 issued the European Norm EN13816:2002 for the definition of quality of service in public passenger transport. According to Section 2.1, we used CEN/TC 320 as the main source of KQIs [2] for the following reasons: (1) It represents a primary tool of PTCs to certify the quality of their routes. (2) It emphasises user-oriented and measurable indicators for monitoring quality, even though it does not specify how to choose a relevant sub-set of indicators. (3) It provides the most relevant and lengthy lists of 8 available service aspects, 30 indicators and 103 sub-indicators. Notably, each service aspect represents the main category of a set of related indicators, which represent the units of measure. Usually, these indicators can be further split into sub-indicators and even sub-sub indicators. For the sake of synthesis, we refer to them all as indicators.

Next, the components and attributes identified in Section 2.2 are used as input elements to guide in the selection process.

3.2   International survey

As described in Section 2.3, we conducted a web-based international survey of experts to obtain their judgements. We selected these experts using two approaches: the former involved academic researchers and the latter practitioners. We selected the academic researchers based on the results of a recent survey of the top public transportation scientists [22] as well as the ranking of the top 50 worldwide universities, as classified by the Centre for World University Rankings and the Academic Ranking of World Universities. Based on these criteria, we selected 291 e-mail addresses from the university websites. Since we found no available worldwide ranking of PTCs, we used a different approach to select a panel of international practitioners. The Union International des Transports Publics (UITP) was our main source, based on the fact that it is a not-for-profit international organisation with 1,400 members from 96 countries worldwide. The UITP provided a list of 615 PTC members, from which we randomly selected practitioners for participation in the survey. We conducted the survey in two steps. First, we asked the experts to perform pairwise comparisons of both methodological and relevance-to-quality components, and related attributes. Their answers were provided according to the scheme shown in Fig. 2. For example, the initial question was "If you were asked to choose a transit service quality indicator based on either its methodological feature or its relevance to service quality, which of the two components would you deem more important to your selection and how strongly so?"

In the second step, the experts graded each indicator against each attribute and assigned them a specific outcome mark according to the schemes, as shown in the example in Fig. 3. Although we used a 1 (worst) to 10 (best) scale in assigning outcome marks, this choice does not affect the generality of the method, which is effective using any scale. Moreover, the similar reliability of different scales from a statistical viewpoint has already been established, although more response options tend to lead to somewhat lower scores [23].

If you were asked to choose a transit service quality parameter based on either its Methodological Feature or its Relevance to service quality which of the two criteria would you deem more important to your selection and how strongly so?

Please mark the appropriate number (1 = equal importance)

9 8 7 6 5 4 3 2 1 2 3 4 5 6 7 8 9

| Methodological Quality (i.e. Mesurability, Ease of Availability, Speed of Availability, Interpretability) | ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ | Relevance to Quality (i.e. Integration of users' and PTC's views, User orientation, Subjective & Objective Measurability, Expressing the amount of passengers) |

In deciding a parameter of 'Transit Service Quality', which of the following attributes would you deem more important for indicator choice and how strongly?

Please mark the appropriate number (1 = equal importance)

9 8 7 6 5 4 3 2 1 2 3 4 5 6 7 8 9

Measurability	○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○	Ease of Availability
Speed of Availability	○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○	Measurability
Interpretability	○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○	Measurability
Speed of Availability	○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○	Ease of Availability
Ease of Availability	○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○	Interpretability
Interpretability	○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○	Speed of Availability
Integration of users' and companies' views	○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○	User Orientation
Subjective & Objective Measurability	○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○	Integration of users' and companies' views
Expressing the amount of passengers	○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○	Integration of users' and companies' views
User Orientation	○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○	Subjective & Objective Measurability
User Orientation	○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○	Expressing the amount of passengers

Figure 2: Step 1 – Pairwise comparisons of components and attributes.

Considering a range from 1-10 please give a mark for every parameter of **External Interface of Public Transport**

	Measurability	Ease of Availability	Speed of Availability	Interpretability	Integration Users-Companies	User Orientation	Subjective & Objective Measurability	Amount of Passengers
to pedestrians	1-10	1-10	1-10	1-10	1-10	1-10	1-10	1-10
to cyclists	1-10	1-10	1-10	1-10	1-10	1-10	1-10	1-10
to taxi users	1-10	1-10	1-10	1-10	1-10	1-10	1-10	1-10
to private car users	1-10	1-10	1-10	1-10	1-10	1-10	1-10	1-10

Considering a range from 1-10 please give a mark for every parameter of **Internal Interface of Public Transport**

	Measurability	Ease of Availability	Speed of Availability	Interpretability	Integration Users-Companies	User Orientation	Subjective & Objective Measurability	Amount of Passengers
entrance/exits	1-10	1-10	1-10	1-10	1-10	1-10	1-10	1-10
internal movement	1-10	1-10	1-10	1-10	1-10	1-10	1-10	1-10
transfer to other PPT modes	1-10	1-10	1-10	1-10	1-10	1-10	1-10	1-10

Figure 3: Step 2 – Classification matrix – An example.

The research design required that two survey waves be conducted in 2017 during January and July by e-mail. In this e-mail, we outlined the aims and scope of the research project and provided guidelines and a link to the web site to access the survey. The first survey step was

completed by 53 experts and the collected data yielded a global consistency ratio of 0.8%. The second survey step was completed by 13 experts. The results are presented below.

3.3 Weight, scoring and selection of indicators

Table 1 reports the derived component weights for the original data, based on the application of the AHP. The results show that experts prefer the "relevance to transit service quality" component.

Table 2 reports the derived attribute weights $h \in H$ and $k \in K$, respectively. We obtained these outcomes by applying the AHP to the original data. The weights in Table 2 illustrate the differences in the opinions of the experts with respect to the methodological features and relevance to transit service quality.

On one hand, these outcomes indicate that experts identify "Interpretability" as the most important attribute to consider when evaluating the methodological feature of a component. On the other hand, "User-Orientation" is the most important attribute to consider when evaluating its relevance to transit service quality. Next, as outlined in Section 2.4, we computed the original $WQIPS_i$s. We conducted experiments 1 and 2, as described in Section 2.5, with B=10,000 and S = 824, respectively. Moreover, for each indicator, we obtained and ranked its $WQIPS_i$ in decreasing order for both the original and stabilised values, as described in Section 2.6. Table 3 shows the outcomes, in which columns 1 to 3 list the three best indicators of each organisational service aspect, as provided by [2], and columns 4 and 5 report the scores of the stabilised and original $WQIPS_i$s.

These results reveal that stabilising the original scores partially changed their expert evaluation rankings (original $WQIPS_i$). Moreover, some relevant changes in the scores of

Table 1: Weights of components.

Component	Symbol	Original weight
Methodological features	$\bar{m}$	0.368
Relevance to transit service quality	$\bar{q}$	0.632

Table 2: Weights of attributes.

Attribute	Symbol	Original weight
Measurability	$\overline{w_1}$	0.2605
Ease of availability	$\overline{w_2}$	0.2854
Speed of availability	$\overline{w_3}$	0.1461
Interpretability	$\overline{w_4}$	0.3079
Integration of users–companies	$\overline{z_1}$	0.1968
User orientation	$\overline{z_2}$	0.3597
Subjective and objective measurability	$\overline{z_3}$	0.2482
Amount of passengers	$\overline{z_4}$	0.1953

Table 3:   Results of stabilised and original *WQIPS$_i$*s for the best three indicators of each service aspect.

Service aspect	Indicator	Sub-indicator	Stabilised WQIPS$_i$	Original WQIPS$_i$
Availability	Operation	Frequency	6.365	6.838
	Dependability		6.202	6.392
	Operation	Vehicle load factor	6.184	6.411
Accessibility	Ticketing availability	Acquisition on network	6.295	6.524
	Ticketing availability	Validation	6.255	6.583
	Internal interface	Transfer to other PPT modes (*)	5.911	5.825
Information	Travel information on normal conditions	About time	6.500	6.788
	General information	About security (*)	6.475	6.812
	Travel information normal conditions	About route	6.454	6.717
Time	Adherence to schedule	Regularity	6.410	6.821
	Adherence to schedule	Punctuality	6.384	6.785
	Length of trip time	In vehicle	6.302	6.419
Customer care	Ticketing options	Payment options	6.295	6.552
	Assistance	For customers needing help	6.285	6.243
	Assistance	At service interruptions (*)	6.204	6.126
Comfort	Seating and persona space	At b/a points	6.230	6.665
	Seating and persona space	In vehicle	6.225	6.649
	Ambient conditions	Cleanliness (*)	6.207	6.471
Security	Freedom from crime	Staff/police presence	6.165	6.369
	Freedom from accident	Avoidance/visibility of hazards	6.046	6.303
	Freedom from crime	Lighting	6.026	6.125
Environmental impact	Pollution	Noise	6.014	6.022
	Pollution	Odour	5.996	6.137
	Pollution	Dust and dirt (*)	5.843	5.747

several indicators appear when comparing the original and stabilised scores. We note that stabilisation tends to reduce the original scores in most cases (see the last column of Table 3). This may be dependent on the distribution of scores obtained for a certain item from all the

experts. Indeed, the distribution is rather homogeneous with a tendency toward central values between the intervals 5 to 7, except for some sporadic high scores whose effect on the mean score is mitigated by the Monte Carlo simulations. Moreover, for indicators marked with "*", stabilisation consistently improves their rankings. Indeed, these indicators are in the top three with respect to the stabilised $WQIPS_i$s even if they had occupied the lowest position in the original ranking. For these cases, it is likely that the original scores are more homogeneous, which yields an increase in the stabilised scores. Furthermore, the stabilisation provides evidence that the three best indicators are those related to information about time, security and route, rather than the frequency, regularity and information about time derived from the original $WQIPS_i$s. These stabilised outcomes are unexpected. They differ from [24], whose authors reported that service reliability (which includes time indicators) is twice as important as frequency and almost seven times more important than information for passengers. In addition, they are derived from a UK practical guideline, which reports passengers' perceptions of local bus services, which is used in rankings of importance.

Finally, we consider the stabilised average scores obtained for each service aspect. Fig. 4 shows plots of the distributions of these average scores for the analysis of the distribution of each individual set of scores and for our evaluation of the most and least important scores. We sorted the boxplots in a top-down manner to provide evidence regarding the most important service aspects, on average. "Time" is the service aspect that yields the highest $WQIPS_i$ on average, whereas environmental impact is ranked as less important. Moreover, Fig. 4 shows the variability of the average scores within each service aspect. In this respect, security is the most homogeneous service aspect as the $WQIPS_i$s of the associated indicators are quite close to the average scores. Conversely, the service aspect "Environmental Impact" shows the most dispersed $WQIPS_i$s.

4 CONCLUSIONS AND RESEARCH PERSPECTIVES

Recent advancements in transit services have captured the interest of experts regarding methods for monitoring the quality of public transport services. Many studies have been proposed for monitoring service quality and appropriate recommendations made. Moreover,
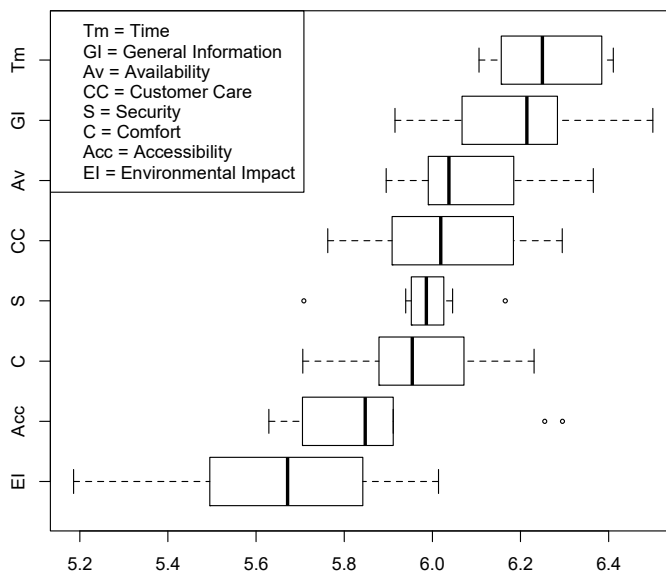

Figure 4: Distribution of average stabilised $WQIPS_i$s for each service aspect [2].

to the best of our knowledge, public transport companies monitor service quality based on ad hoc indicators rather than on those derived using a clear and well-organised methodology.

However, the choice of specific KQIs represents a crucial element. This paper contributes to transit service research by proposing a robust methodology for deriving top KQIs for describing transit service quality. The methodology comprises six steps. First, a lengthy list of indicators is identified. Next, the goal as well as the components and attributes are defined for each indicator, and leading experts are involved to elicit judgement for each indicator.

Next, each indicator is evaluated and the most promising set is determined. We demonstrated the robustness of this method twice. First, we gathered data input comprising indicator weights and marks in a web-based international survey. Secondly, we stabilised the outcomes using Monte Carlo simulation methods.

This method is anticipated to advance the state of the art both theoretically and in practice. It fills a gap in the characterisation of transit quality (i.e., the selection of indicators). Moreover, it provides outcomes, which can be readily applied by every PTC to monitor its service quality.

To the best of our knowledge, this is the first paper providing a set of KQI that describe service quality based on a robust methodology.

To conclude, we suggest two future research directions. First, this method can be easily applied to other transportation modes (e.g., air, maritime and so on) as well as other general services. Second, a more challenging study based on social web and advancements in web semantics could involve users rather than experts to construct a list of indicators. This is made possible by the use of feedback obtained from users on social media regarding their experiences with transportation companies (e.g., Facebook, Twitter, Instagram). In this way, new indicators can be derived or older ones confirmed.

REFERENCES
[1] Le Callet, P., Möller, S. & Perkis, A. (eds), Qualinet White Paper on Definitions of Quality of Experience. European network on quality of experience in multimedia systems and services (COST Action IC 1003), Lausanne, Switzerland, Version 1.2, March 2013.
[2] Transportation–Logistics and services, European Standard EN 13816, 2002.
[3] De Oña, J. & De Oña R., Quality of service in public transport based on customer satisfaction surveys: A review and assessment of methodological approaches. *Transportation Science*, **49**(3), pp. 605–622, 2014. DOI: 10.1287/trsc.2014.0544.
[4] de Oña, J., De Oña, R., Eboli, L. & Mazzulla, G., Index numbers for monitoring transit service quality. *Transportation Research Part A: Policy and Practice*, **84**, pp. 18–30, 2016. DOI: 10.1016/j.tra.2015.05.018.
[5] Eboli, L. & Mazzulla, G., A methodology for evaluating transit service quality based on subjective and objective measures from the passenger's point of view. *Transport Policy*, **18**(1), pp.172–181, 2011. DOI: 10.1016/j.tranpol.2010.07.007.
[6] Eboli, L. & Mazzulla, G., Performance indicators for an objective measure of public transport service quality. *European Transport*, **51**, pp. 1–21, 2012.
[7] Nathanail, E., Measuring the quality of service for passengers on the Hellenic railways. *Transportation Research Part A: Policy and Practice*, **42**(1), pp.48–66, 2008. DOI: 10.1016/j.tra.2007.06.006.

[8]   Tyrinopoulos, Y. & Antoniou, C., A complete methodology for the quality control of passenger services in the public transport business. *European Transport*, **15**, pp. 1–16, 2008.

[9]   Tyrinopoulos, Y. & Antoniou, C., Public transit user satisfaction: variability and policy implications. *Transport Policy*, **15**(4), pp. 260–272, 2008.
DOI: 10.1016/j.tranpol.2008.06.002.

[10]   dell'Olio, L., Ibeas, A. & Cecin, P., The quality of service desired by public transport users. *Transport Policy*, **18**(1), pp. 217–227, 2011.
DOI: 10.1016/j.tranpol.2010.08.005.

[11]   Hassan, M.N., Hawas, J.E. & Ahmed, K., A multi-dimensional framework for evaluating the transit service performance. *Transportation Research Part A: Policy and Practice*, **50**, pp. 47–61, 2013. DOI: 10.1016/j.tra.2013.01.041.

[12]   Stradling, S., Anable, J. & Carreno, M., Performance, importance and user disgruntlement: A six method for measuring satisfaction with travel modes. *Transportation Research Part A: Policy and Practice*, **41**(1), pp. 98–106, 2007. DOI: 10.1016/j.tra.2006.05.013.

[13]   Chen, X., Yu, L., Zhang, Y. & Gou, J., Analyzing urban bus service reliability at the stop, route, and network levels. *Transportation Research Part A: Policy and Practice*, **43**(8), pp. 722–734, 2009. DOI: 10.1016/j.tra.2009.07.006.

[14]   Lin, J., Wang, M.L. & Barnum, P.D., A quality control framework for bus schedule reliability. *Transportation Research Part E: Logistics and Transportation Review*, **44**(6), pp. 1086–1098, 2008. DOI: 10.1016/j.tre.2007.10.002.

[15]   Barabino, B. & Di Francesco, M., Characterizing, measuring, and managing transit service quality. *Journal of Advanced Transportation*, **50**(5), pp. 818–840, 2016. DOI: 10.1002/atr.1377.

[16]   Mahmoud, M.M., Hine, J. & Kashyap, A., Bus transit service quality monitoring in UK: A methodological framework. *Proceedings of the ITRN*, 2011.

[17]   Castillo, H. & Pitfield, D.E., ELASTIC—A methodological framework for identifying and selecting sustainable transport indicators. *Transportation Research Part D: Transport and Environment*, **15**(4), pp. 179–188, 2010.
DOI: 10.1016/j.trd.2009.09.002.

[18]   Ibeas, A., dell'Olio, L. & Montequín, R.B., Citizen involvement in promoting sustainable mobility. *Journal of Transport Geography,* **19**(4), pp. 475–487, 2011.
DOI: 10.1016/j.jtrangeo.2010.01.005.

[19]   Zelany, M., A concept of compromise solutions and the method of the displaced ideal. *Computers &Operations Research*, 1(3–4), pp. 479–496, 1974. DOI: 10.1016/0305-0548(74)90064-1.

[20]   Saaty, T.L., *The Analytic Hierarchy Process,* McGraw Hill: New York, 1980.

[21]   Saaty, R.W., The analytic hierarchy process—what it is and how it is used. *Mathematical modelling*, 9(3–5), pp. 161–176, 1987.

[22]   Heilig, L. & Voß, S., A scientometric analysis of public transport research. *Journal of Public Transportation*, **18**(2), pp. 111–141, 2015. DOI: 10.5038/2375-0901.18.2.8.

[23]   Dawes, J., Do data characteristics change according to the number of scale points used? An Experiment Using 5-Point, 7-Point and 10-Point Scales. *International Journal of Market Research*, **50**(1), pp. 61–104, 2008. DOI: 10.1177/147078530805000106.

[24]   Balcome, R. et al., The demand for public transport: a practical guide. TRL Report, TRL 593, TRL Limited, 2004.