Transportation Research Part A 157 (2022) 258–273

Contents lists available at ScienceDirect

Transportation Research Part A

journal homepage: www.elsevier.com/locate/tra

Is higher quality always costly? Marginal costs of quality: Theory and application to railway punctuality

Andrew S.J. Smith *, Manuel Ojeda Cabral

Institute for Transport Studies, University of Leeds, UK

ARTICLE INFO

Keywords:
Cost of quality
Delay costs
Proactive costs
Reactive costs
Reliability
Marginal cost
Railway
Efficiency

ABSTRACT

In the railway sector, there has been much discussion about the costs of delays to passengers and their willingness-to-pay to reduce them. However, the cost of delays in the supply side of transport markets have received far less attention (Van Oort, 2016). This paper fills several gaps in the transport and railway literature by studying the relationship between the costs of railway supply and travel time reliability. First we articulate a generic theoretical framework for the relationship between costs and quality in railways, building on past contributions in other sectors and bringing together diverse and currently disconnected literatures. A key new element of the framework is the explicit introduction of the concepts of 'marginal proactive cost' and 'marginal reactive cost'. Second, through the lens of this framework, we empirically study the relationship between the cost of passenger train operation companies (TOCs) in Great Britain (2011 to 2015) and the reliability of their services, thus producing the first estimates in the literature of the elasticity of train operating company cost with respect to train delays, and in turn marginal cost of reducing delays. We find that for most TOCs improving quality is costly but, for some, quality improvements would be associated with lower costs overall, indicating that some operators are on a sub-optimal portion of the cost-quality curve. The framework and analysis can be used to aid quality related decisions in the railways including the design of incentive regimes; and can also be applied to other cost-quality contexts, in and outside transportation.

1. Introduction

The reliability of a transport network is an important element of the market, both from the supply and demand perspective. Reliability, here understood as the degree of certainty of travel times[1], has been mainly studied from the perspective of travel demand. Unreliability is costly and unpleasant to transport users and hence affects travel demand negatively. An extensive body of literature has studied how passengers value reliability (see Carrion and Levinson, 2012) and how changes in reliability affect demand (see Wardman and Batley, 2014). As with the high costs to users, it is reasonable to assume that improvements in reliability are costly to produce, and indeed cost-reliability relationships in transport and utilities more widely are at the heart of the relationship between regulated firms and regulatory bodies / funders. However, empirical evidence on the relationship between reliability and the supply side of the transport market is scarce. A limited number of studies have been identified in this area and none provide a theoretical discussion of the

* Corresponding author at: Room 2.13, Institute for Transport Studies, 34-40 University Road, University of Leeds, LS2 9JT, UK.
E-mail address: a.s.j.smith@its.leeds.ac.uk (A.S.J. Smith).

[1] Reliability is, technically, a broader concept that could be defined as the certainty of the service delivery, including travel times (in-vehicle time, waiting time, etc.), seat availability and other quality aspects (see Van Oort, 2016). Here we focus on the reliability of travel times.

https://doi.org/10.1016/j.tra.2022.01.007
Received 12 March 2020; Received in revised form 6 December 2021; Accepted 13 January 2022
Available online 16 February 2022

mechanisms through which cost and reliability may be related.

It is therefore necessary to increase our understanding of the cost-reliability relationship. Nonetheless, we shall see that this relationship is not straightforward. In regulatory contexts, it is generally assumed that if firms or regulators wish to improve quality, these improvements should come at a cost. That is, it is expected that the marginal cost of quality should be positive, such that if higher quality is required, more money needs to be spent; or conversely that money can be saved if a lower quality standard can be accepted.

However, there is some evidence in the literature of a negative relationship between cost and quality, hence indicating that poor quality can be associated with higher costs (e.g. in the water, health and airline sectors; see CEPA, 2014; Gutacker et al., 2013; NEXTOR, 2010). Such a relationship indicates a sub-optimal position since costs could be reduced and quality improved simultaneously. This raises questions as to how regulators or governments (funders) should view the relationship between costs and quality and set targets and funding allowances accordingly. In some instances, a negative cost-quality relationship can create perverse incentives in economic regulation as it could imply giving companies an increased cost allowance for delivering poor quality. In the UK system of economic regulation, for example, which is applied to multiple network industries (see, for example, Ofwat, 2019; Ofgem, 2021 for the water and energy sectors) - regulators determine a cost allowance based on the estimated relationship between cost and the included variables in the model; therefore, a negative sign on the quality variable would imply giving lower cost allowances for higher quality performance. Similar regulatory regimes apply in other countries for example, Germany, Japan, Australia and Ireland[2] (see for example, Australian Competition and Consumer Commission, 2012; Sumicsid, 2007; Lawrence et al., 2014).

Considering the wider efficiency / productivity estimation literature, it is also important to control for quality in order to produce credible efficiency comparisons. Low cost operations can always be achieved; however, these may be delivered at the expense of quality (see e.g. Hart et al., 1997). However, in general the extensive literature in this area tends to neglect quality, partly because it is hard to obtain reliable and comparable data; and partly because theoretically the relationship has not been fully formalized.

This papers builds upon the existing railway and transport literature, while drawing ideas from other literatures. The contributions of the paper are as follows. First, we articulate a framework that provides a theoretical foundation for the mechanisms through which to understand the marginal cost of improving quality in railways, building on past contributions in other sectors and bringing together diverse and currently disconnected literatures. In addition, a key new element of the framework is the explicit introduction of the concepts of 'marginal proactive cost' and 'marginal reactive cost'. The gaps in the literature make it necessary to enhance our understanding and formalise the relationship between cost and quality in railways, in particular to grasp why a negative coefficient (as opposed to the expected positive sign) may emerge between cost and quality (because of high reactive costs) and to identify situations in which rail firms may be operating in a sub-optimal way (operating with both high costs and poor quality). In turn this enables "win–win" situations to be identified whereby both costs can be reduced and quality enhanced at the same time.

Second, through the lens of this framework, we empirically study the relationship between the cost of passenger train operation companies (TOCs) in Great Britain and the reliability of their services. The paper is therefore the first to focus on the direct theoretical relationship between cost and reliability in passenger rail operations and the estimation of elasticities and marginal costs of reliability improvements.

Along with Sweden, Great Britain has gone furthest in Europe in terms of introducing competition in passenger rail services and also has the most advanced system of compensation for unreliability (train operator to train operator and operator to infrastructure manager) anywhere in the world. Understanding the marginal cost of quality is potentially important in calibrating such a compensation regime, but such understanding is currently vague. Although the main focus of this paper is on TOCs, the framework provided is generic and can be applied to the context of railway infrastructure managers, other transport modes and, more broadly, other areas where quality is relevant.

The paper is structured as follows. Section 2 contains the literature review, before the presentation of the theoretical framework for the cost-quality relationship in Section 3. Section 4 then discusses how this theoretical framework can be understood in the specific context of railways, as an introduction to the empirical application (Great Britain's railways). The data and methodology are set out in Section 5, before the results in Section 6. Section 7 concludes.

2. Literature review

2.1. Railway context

Affuso et al. (2002) measured the efficiency of British TOCs soon after privatisation took place, and related the estimated efficiency scores to levels of punctuality. Their findings revealed a negative correlation between efficiency and punctuality, i.e. improved punctuality would come at the cost of reduced productive efficiency. However, the link was statistically very weak. Kennedy and Smith (2004), used a distance function approach that included delays as an input alongside costs to model the efficiency of Britain's infrastructure manager – as a side result, that study found a weakly significant negative relationship between delays (caused by Network Rail) and infrastructure costs – that is, low delays (high quality) is associated with high costs.

Abate et al. (2013) conducted a European study of the relationship between reliability and productivity in rail. They describe various mechanisms that link punctuality and productivity and take the view that the overall relationship is ambiguous due to the presence of countervailing forces. For instance, the "effort effect" mechanism (punctuality can mean sacrificing output productivity),

[2] Note this list is not exhaustive but includes countries where econometric cost function estimation has been used as part of the benchmarking analysis underpinning cost allowances.

and the "demand effect" mechanism (punctuality attracts passengers and hence can mean increased productivity). However, they do not formalize the relationship prior to their empirical analysis. They conclude that improving reliability is not necessarily linked to reductions in productivity (importantly, noting that the measure of productivity itself only considers inputs and physical outputs, but neglects quality). They interpret the empirical results as evidence that the different effects compensate each other in practice. More recently, Link (2019) studies the impact of including quality measures on efficiency scores, finding evidence that the inclusion of punctuality altered the efficiency rankings of operators; some operators with higher punctuality ranked better once punctuality was taken into account.

Estache et al. (2007) dealt with the potential trade-off between quality and output in the Brazilian freight railway. They find evidence of the existence of such a trade-off in some periods but also evidence that quality can be positively correlated with productivity depending on the regulation scheme, which marries well with the countervailing effects discussed by Abate et al. (2013). However, quality was defined as an index combining safety and speed of the transported freight, and did not include punctuality. More recently, Van Oort (2016), within an illustration of how to introduce reliability improvements in cost-benefit analyses for bus projects, provides a brief discussion of the implications of unreliability for a transport operator. In the discussion, Van Oort (2016) intuitively considers that enhanced reliability is associated with lower (not higher) costs for a transport operator, but does not provide details of the rationale for this statement.

Typically the literature does not explore the primary link between reliability and (observable) costs, but rather the links between quality and productivity/efficiency. Furthermore, the data used might have not helped to uncover the relationships. The heterogeneity among the several European Railways analysed in Abate et al. (2013) and the difficulties to obtain and combine data from the different countries might have been a drawback. Affuso et al. (2002), on the other hand, had limited data just after the British railway privatization and their measure of punctuality was a measure that also included the performance of the infrastructure manager (whose inputs were not included). Finally, the research questions in Kennedy and Smith (2004) are closer to ours, but they focused on the infrastructure manager; further, they did not explicitly discuss the underlying theoretical relationship between quality and costs.

2.2. Other contexts: Outside the railway

NEXTOR (2010) analyses the cost of delays to airlines, passengers and more broadly to the US economy. Their study from the airlines perspective takes the supply angle that we are interested in. They rationalize what the costs of delays are to the airlines. They find a significant positive effect of the combination of delay and buffer time on airline costs, and use the model results to derive estimates of the total cost of delay time to airlines for a given year. However, their theoretical framework does not grasp the underlying links between the different elements involved, and hence it is not clear how delays relate to costs. Peterson et al. (2013) carried out a new analysis of the overall cost of airline delays to the economy using a general equilibrium model. These works are more concerned with 'how costly are delays', and less with 'how costly are delay improvements'. Other air-transport studies focus on efficiency and performance (Assaf et al., 2014; Merkert et al., 2015).

The economics and management literature revolves around the seminal work of Juran (1951), Feigenbaum (1956) and Masser (1957). Juran (1951) introduced the concept of "quality costs", pointing out that companies can face failure costs due to poor quality and therefore there is a benefit in avoiding poor quality. Feigenbaum (1956) and Masser (1957) extended the definitions of cost types, leading to what is known as the Prevention Appraisal Failure (PAF) model, a type of Cost of Quality Model (COQ). This and other models that relate costs and quality were developed later (e.g. Crosby, 1979, Juran and Gryna, 1988). They categorized the costs associated with quality into three categories: prevention, appraisal and failure costs.

Juran et al. (1962) first discussed the possibility of trade-offs between prevention and failure costs. The developed Cost of Quality Model (COQ) suggests a quadratic relationship between costs and quality (see Fig. 1). The COQ model depicted in Fig. 1 (PAF model) is

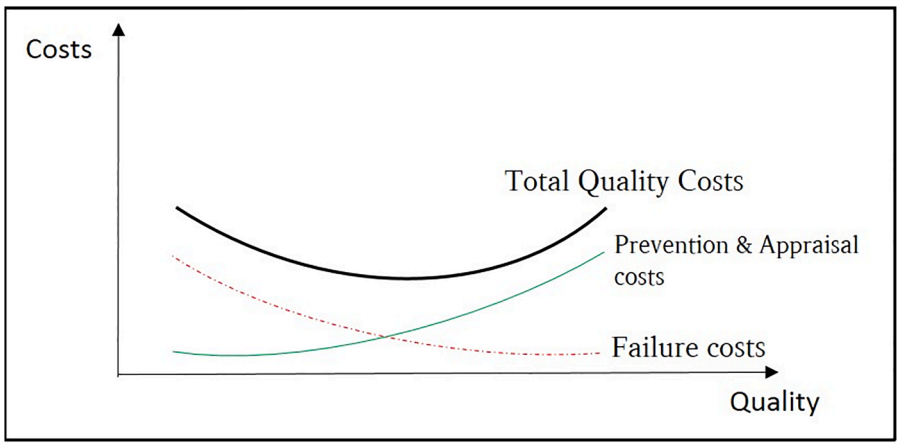

Fig. 1. Illustration of a classical cost of quality model. Source: adapted based on various works (e.g. Juran et al., 1962; Schiffauerova and Thomson, 2006).

the most widely used in this area (see Schiffauerova and Thomson, 2006 for recent review of this literature). The rationale for the cost curves are that increased prevention and appraisal costs are needed if higher quality is to be achieved; however, failure (low quality) also creates costs, thus this aspect of costs falls with quality. The two countervailing pressures intersect at some point thus creating a U-shaped overall cost-quality curve. For reviews on this literature see Hwang and Aspinwall (1996) and Schiffauerova and Thomson (2006). The literature also includes empirical studies that have attempted to calculate the costs of quality (e.g. Freeman, 1995; Srivastava, 2008; Peimbert-García et al., 2016) in different manufacturing and service industries. However, the aforementioned theoretical and empirical works did not focus on the analysis of the somewhat more subtle economic concepts of the elasticity of cost with respect to quality and the marginal cost of quality. Instead, they have looked at total costs associated with quality and used accounting, management and simulation methods to unpick the different types of quality costs (e.g. appraisal, prevention and failure costs).

The literature also contains examples of theoretical frameworks for the relationship between cost and quality combined with empirical applications in other utility sectors. Interestingly, these applications provide frameworks that can be seen as equivalent to the wider Cost of Quality Models (COQ). However, they all seem unaware of the existence of this more generic body of literature. In the energy sector, Jamasb et al. (2012) and Coelli et al. (2013) develop two econometric approaches to estimate the marginal costs of a quality enhancement for UK electricity distribution firms. Jamasb et al. (2012) also provides a theoretical framework emphasising the need to understand the different types of costs associated with quality, which they refer to as preventative and reactive costs. This implies that the direction of the overall relationship between cost and quality is not straightforward, but this is not linked to existing COQ models (even though the concepts of prevention and reactive costs are clearly similar to the concepts of prevention & appraisal costs and failure costs in the earlier framework). They also show the importance of temporal dynamics in the quality-cost relationship, recognising that part of the relationship may not be contemporaneous in the electricity context.

In the field of health economics, Gutacker et al. (2013) hypothesise that the marginal costs of quality are non-constant and cost and quality may be explained by a U-shaped relationship. The explanation is that, although high quality is linked to high costs, low quality may also be linked to high costs under some circumstances. For example, where hospitals have not implemented subtle quality improvements that also bring cost savings, such as early mobilization of patients after joint operations; thus leading to complications that require further treatment and higher costs (such that costs are high and the patient experience is low). However, again there is no link with generic COQ literature and no formalization of the model.

In summary, both the wider literature on quality and costs and the specific quality studies in different utility sectors contribute excellent ideas. There are however gains to be exploited by bringing together these diverse bodies of research that so far have remained disconnected – and in developing a more comprehensive and generic framework for the relationship between cost and quality, where the marginal cost of quality is at the centre of the discussion. Our framework will be applied to the study of railway reliability, but is general enough to be applied in a wide range of cost-quality contexts. It should be noted that we have focused on the literature concentrating on the relationship between quality and costs. Future research in this area will also benefit from incorporating insights from other relevant bodies of literature, such as those concerned with the understanding and development of quality indicators (e.g. Cascetta and Cartenì, 2014) or consumers' perception of quality (e.g. dell'Olio et al., 2011; Mayat and Wheat, 2020) and its impact on travel demand (see Wardman and Batley, 2021, for a recent review of the demand impacts of train punctuality). Customer satisfaction can be seen as a gap between what customers desire and experience in practice (see for example Stradling et al., 2007), and the experienced levels of punctuality impact upon users' travel choices (Wardman and Batley, 2021). Whilst our paper utilises train based delay measures (see section 4 for further details), this does not necessarily correlate directly with passengers' perception of lateness or more widely with passenger satisfaction. For that reason, passenger satisfaction measures have also been incorporated into rail franchise agreements in Great Britain (see the Williams Rail Review, 2019). The issue of how to incorporate such measures into cost functions – and interpret the results in respect of elasticities and also performance management, given that rail firms do not directly control satisfaction – could be a useful direction for further research.

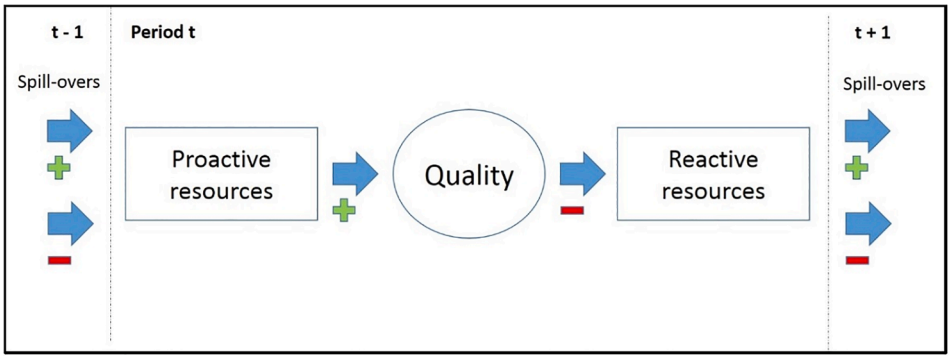

Fig. 2. Conceptual framework: Quality and use of resources. Note that in Fig. 2 the arrows reflect the direction of relationship such that, for example, if proactive resources are committed, this will impact on quality. The positive sign indicates that increased proactive resources will lead to higher quality; whilst the negative sign indicates the high quality performance results in lower reactive resources being committed. Likewise the positive and negative signs on the left and right hand side of the diagram indicate that the relationships between costs and quality that spill over into different time periods (e.g. t-1 or t + 1) could be positive or negative.

3. The relationship between cost and quality: General theoretical framework

In this section we set out and develop a generic theoretical framework for the cost-quality relationship which explicitly addresses the cost-quality relationship in terms of marginal costs – and which could be applied to any industry. Section 4 then discusses how this theoretical framework can be understood in the specific context of railways.

The theoretical framework can be derived from the classical COQ model known as PAF model. It also brings together the existing literature from different utility sectors with the wider economics and management literature; specifically building on the frameworks of Juran et al. (1962) in quality research, Jamasb et al. (2012) in the electricity sector and Gutacker et al. (2013) in the health sector.

We first introduce the framework conceptually, as a generalization of Jamasb et al. (2012)'s model (Fig. 2). The key element of the relationship is the need to recognise two types of cost: proactive costs and reactive costs. This categorization is the common factor across all bodies of literature. Proactive costs are those related to the proactive use of resources aimed at enhancing quality (e.g. prevention and appraisal costs). Part of the literature (e.g. Jamasb et al., 2012, Yu et al., 2009), refer to these as preventative costs. However, this places the emphasis on preventing something going wrong with quality. Importantly, the term proactive more generally accommodates any use of resources aimed at improving quality.

Proactive costs hold a positive relationship with quality. That is, in order to enhance the quality of service provision, additional resources need to be committed. In general these resources may involve a higher quality of capital input, or increased staffing in order to increase preventative maintenance activity. We discuss specific examples in respect of railways in section 4 below. On the other hand, reactive costs arise from the use of resources reactively after a given quality has been produced or delivered (e.g. failure costs). Reactive costs are incurred to correct poor quality and/or to compensate the affected parties (e.g. consumers). Such costs could be high, since regulatory targets will require service to be resumed quickly in the case of failure (e.g. a blackout for an energy network), and resources will therefore have to be activated quickly to bring service back into operation. Potentially compensation to consumers could also be high. We discuss specific examples in respect of railways in section 4 below. Therefore, reactive costs hold a negative relationship with quality. To summarise, a decision-maker can spend money to improve quality and, the higher the quality, the lower the cost of the consequences associated with quality failures.

Additionally, the framework recognises that the two-directional relationships are not restricted to a particular time frame. For a given time period t, both relationships between costs and quality may realize within the time period t or extend beyond it. For example, proactive costs by electricity firms in year t to improve quality may only translate into quality improvements in year $t+1$ (e.g. Jamasb et al., 2012). Reactive costs in the form of compensations to customers will probably be contemporaneous for many firms. This temporal dimension is recognised in the framework through the concept of spill-overs beyond period t. The extent and importance of the temporal dimension will highly depend on the context of study.

It is also important to specify who the decision-maker is. For example, we can take the perspective of a private firm or the perspective of a regulator interested in social welfare. The former will focus on the costs incurred by the firm, whereas the later will also include any additional costs incurred by other agents (e.g. customers).

For economic analysis, the framework needs to be formalized in terms of marginal costs. The distinction between proactive and reactive costs allows us to separate, theoretically, the marginal cost of quality (MCq) in two components: the marginal proactive cost of quality (MPCq), and the marginal reactive cost of quality (MRCq), such that:

$$MCq = MPCq + MRCq \tag{1}$$

where:

$$MPCq > 0$$

$$MRCq < 0$$

The signs of the two elements of marginal cost were explained earlier in this section.

To the best of our knowledge, this explicit distinction of marginal (proactive and reactive) costs is a novelty of our framework. A unit increase in quality is associated with: i) a positive change in proactive cost (i.e. $MPCq > 0$), ii) a negative change in reactive costs ($MRCq < 0$). $MPCq > 0$ reflects that investments in quality are costly, and $MRCq < 0$ reflects that there are savings to be made (through avoiding costly consequences) by improving quality. Consequently, a unit increase in quality is associated with a change in total cost (MCq) that can be positive, zero or negative. The magnitude of MCq will depend on the magnitudes of MPCq and MRCq. Fig. 3 below provides a graphical representation of the theoretical framework.

Theoretically, we can also assume that the magnitudes of MPCq and MRCq are interrelated and depend on the current level of quality. The higher the quality, further improvements become more expensive: MPCq increases with quality:

$$\frac{\partial MPCq}{\partial q} > 0$$

Similarly, the higher the quality, the scope for savings from reductions of reactive costs decreases: MRCq decreases, in absolute terms, with quality.

$$\frac{\partial |MRCq|}{\partial q} < 0$$

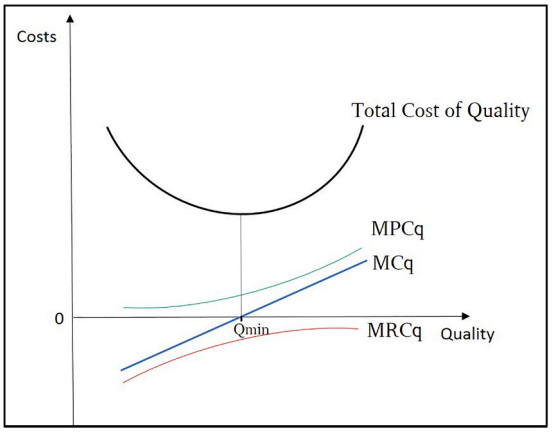

Fig. 3. Costs and quality theoretical framework, from a marginal cost perspective.

Since MRCq is always negative and becomes smaller in absolute terms as quality increases, the overall relationship between MCq and quality is as expected according to economic theory: the marginal cost of quality (MCq) increases with quality, i.e. $\partial MCq/\partial q > 0$.

However, strict economic theory dictates that MCq should always be positive. As we shall see, the possibility of a negative MCq exists but would be associated with inefficient choices (i.e. departures from rational behaviour). Economically efficient decision-makers should indeed face a positive MCq. This framework thus reconciles COQ models and the traditional economic theory behind cost functions. Graphically, this framework implies a U-Shaped cost curve, in line with most COQ models (e.g. Juran et al., 1962; Schiffauerova and Thomson, 2006) and some empirical applications (e.g. Gutacker et al., 2013); see Fig. 3.

There exists a point (Qmin) which determines the minimum level of quality that a decision-maker should produce (e.g. Juran et al., 1962). Below that point, letting quality worsen also brings an increase in cost due to high reactive costs (MPCq < -MCRq, hence MCq < 0). In other words, to the left of Qmin improving quality brings net cost savings which more than compensate the proactive resources committed. On the other hand, quality levels greater than Qmin can only be achieved at an extra cost: MCPq > -MCRq, hence MCq > 0. In other words, to the right of Qmin the reactive cost savings do not compensate for the increased proactive costs, and the quality improvement does have a net positive cost. The decision-maker should therefore consider the marginal benefit of the increase in quality to judge whether the positive associated MCq is worth the commitment to higher proactive costs[3]. While optimality analysis requires also consideration of demand, the framework does show that: under the decision-making assumptions of neo-classical economic theory, we should only find firms producing output with at least a level of quality Qmin, such that:

$$MPCq \geq -MRCq$$

$$MCq \geq 0$$

This does not, however, prevent us from observing negative MCq in some contexts, as the existing literature has shown, but importantly such a case would imply sub-optimal behaviour.

To summarise, the framework provides researchers with a theoretical guidance to approach the study of marginal costs of quality in any particular context. In regulated contexts, it provides guidance as to how regulators / policy makers should set quality targets / incentivise quality improvements. It has been built up from previous contributions, drawing together a wide and previously disconnected set of literatures – also with the addition of making the explicit distinction between marginal (proactive and reactive) costs, which, to the best of our knowledge, is novel.

4. The relationship between cost and quality: Positioning the theoretical framework in railways

Having articulated the general theoretical framework, this section discusses how the framework can be understood in railways, particularly focusing on the British railway sector which is our empirical application.

In Europe, and Great Britain, train delays and cancellations are seen as a major problem. Efforts have been made in the last decade to improve the reliability of the services, but the figures from the industry indicators of reliability (e.g. the Public Performance Measure, PPM[4]) and customers' satisfaction surveys indicate there is still a long way to go. Fig. 4 below shows how train punctuality, measured by PPM, has stagnated over the last decade up to the period covered by our data. Even punctuality levels of approximately 90% on average is not viewed as a sufficiently high level of performance overall (see for example, House of Commons Transport Committee, 2017, which discusses "continued network underperformance"). Further, these figures often hide substantial variations

[3] Here we are holding output constant – i.e. only cost and quality are allowed to vary.
[4] This measure combines performance in terms of punctuality and cancellations.

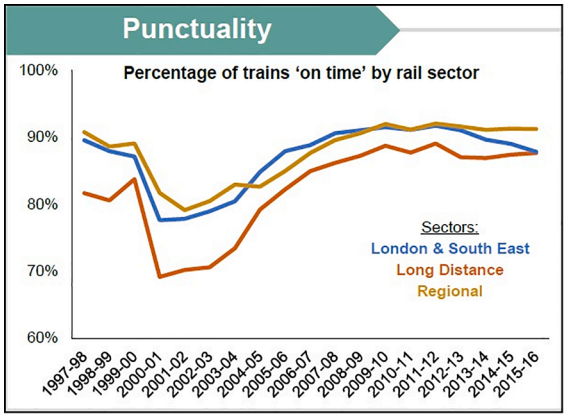

Fig. 4. Rail punctuality in the UK. Source: DfT (2017), Rail Trends Factsheet.

across different parts of the network, omit very small delays and do not capture the extent/severity of delays. Here it should be noted that very high punctuality performance can be achieved in some railway contexts, most notably Japan (see for example, Rail Delivery Group, 2018).

The British Rail Technical Strategy (see TSLG, 2012) set out to radically improve the "4Cs" (increase capacity, reduce carbon, lower cost and improve customer satisfaction). This focus was re-iterated in the update to the strategy in 2021, with a focus on five priorities, including "improving reliability". The problem of how to ensure good performance – and enhance performance - whilst also growing traffic is a key challenge in the UK but also more widely in Europe. For example, the European rail research effort is targeting technology developments that will achieve a 50% increase in reliability (see Shift2Rail, 2015). In the UK, in passenger surveys, rail passengers repeatedly identify delay performance the most important factor driving their satisfaction in using the railways (see Transport Focus, 2020).

While there is evidence of how much travellers are willing to pay for reliability improvements in the UK (e.g. ARUP et al., 2015), little is known about the cost of improving reliability from the supply perspective. The British railway is a vertically separated industry, in line with some other European countries such as Sweden. Infrastructure management is the responsibility of a legally separate and state owned company, Network Rail, whilst passenger operations are delivered primarily by twenty Train Operating Companies (TOCs) with the monopoly rights to run services on given routes allocated by competitive tendering. In some cases the competitive tender selection process is based on which operator can deliver the service specification for the lowest possibly subsidy; on profitable franchises, operators compete based on their willingness to pay the highest premium to government. An independent regulator, the Office of Rail and Road (ORR), is in place to promote an adequate functioning of the industry. It is well known that vertical separation poses risk on reliability levels and there is a need for performance incentives for both the infrastructure manager (IM) and TOCs.

Quality performance in various dimensions – including delay minutes – is embedded within the franchise agreements signed between the operator and the relevant national, or regional franchising authority. Benchmarks are set and if performance falls short, operators are required to produce a recovery plan and may also incur a penalty. In extreme situations poor performance could lead to the franchise agreement being terminated early.

In addition to the targets set out in the franchise agreements, Britain has a sophisticated performance incentives system called Schedule 8 (S8); see Network Rail (2016). S8 works around a set of targets customised by TOC and type of service. IM and TOCs must reward (penalize) each other if punctuality is better (worse) than an agreed target. However, how costly is it for the IM and TOCs to improve reliability? And, is it always costly? In other words, could a company improve reliability and save costs overall? An answer to these questions would seem necessary for optimised decision-making in the industry and for the design of any performance incentives system. Yet, it is far from clear whether we know the answer; which is the focus of this paper.

We discuss the implications of how quality is determined / targeted for the econometric model (recognising the potential endogeneity) in section 5.

The quality aspect of our application is the 'reliability of travel times'. Travel time reliability can be measured in several ways: e.g. number of minutes of lateness or delays or the degree of punctuality (proportion of services on time). These three terms (namely lateness, delays and (un)punctuality all refer to the more generic concept of (un)reliability. Our paper employed the generic term reliability for the theoretical discussion and uses a metric of delay minutes for the applied work.

The context of our empirical application is railway operations in Great Britain and the decision-making units are the TOCs. We use a panel of eighteen train operating companies over a 5-year period (2011 to 2015). However, in vertically separated industries, delay minutes can be attributed to more than one decision-maker: e.g. the infrastructure manager or multiple operating companies. This is a problem that needs to be overcome, since we can only infer the marginal cost of improving quality if the decision-making unit is responsible for it. Conveniently, the railway industry in Great Britain allocates every minute of delay recorded across the IM and TOCs based on their responsibility. In 2016, the infrastructure manager (IM; Network Rail) was deemed responsible for approximately 60% of delays, with the TOCs being responsible for the remaining 40% of observed delay minutes. Our application will focus on delay minutes caused by TOCs as the measure of quality / reliability.

Examples of proactive costs, where a TOC can intervene / "invest" with the aim of reducing delay minutes, include, inter alia: the inspection and maintenance regime with respect to the rolling stock (preventive investment to prevent failures), choice of rolling stock and fleet size (e.g. vehicles fleet in case a failure occurs); and staff (e.g. railway operations, timetabling and general management).

Reactive costs are present in the railway for one main reason: reliability embodies a commitment and strong customers' expectations. Reliability is not simply quality of service like other aspects such as the type of seat or the vibration experienced on the train (i.e. comfort). Reliability is a "promised quality": the tickets bought by costumers are a contract between them and the train company, which includes the price and the travel time advertised in the schedule. The notion of 'promised quality' implies there will be reactive costs when quality is not achieved – and these will be incurred rather quickly with each delay.

First, since TOCs must run scheduled trains anyway, delays (i.e. low quality) mean additional running time for train services that inevitably translate into increased operational costs, including staff extra hours and energy costs. Secondly, the reactive costs are even larger when affected customers are entitled to compensation from significant delays and cancellations. In the British context, passengers are entitled to monetary compensation in the event of delayed trains, and this aspect of compensation is part of the reactive costs of delays and is included within our measure of operator cost. Additionally, there will be interactions between demand and reliability which will not be reflected in TOC costs but instead in TOC revenues. For example, high delays can mean foregone revenue, which is a reactive cost which could be saved by investing in reducing delays. We limit our analysis to direct observable costs for the purpose of the study, but the framework can also deal with a more holistic analysis that takes additional (non-observable) costs such as revenue impacts into account.

All TOCs should naturally have the incentives to run services such that the MC of reducing delays is positive (i.e. locate above Q_{min}, as depicted in Fig. 3). Additionally, incentives to invest in quality beyond this minimum point would depend on demand and the marginal benefits of reliability improvements (including incentives provided by any regulatory performance regime). As argued by Gutacker et al. (2013), the judgement of the regulator on incentives should differ depending on where the operators are in the cost-reliability curve. Increases in cost would not always be justified by potential increases in reliability, since increases in reliability do not necessarily have to come at an extra cost (if firms produce quality below Qmin). However, also, because there are natural gains in moving beyond Qmin, regulators should also address and remove any industry constraints in order to prevent TOCs from operating where quality $<$ Qmin.

In the context of railway operation, some reactive costs will happen immediately (at least, within-year) if reliability is low: e.g. staff extra hours and compensation paid to passengers. Positive quality effects will also be contemporaneous for some proactive costs such as hiring staff to enhance the rolling stock maintenance regime, or improving operations. Hence, we hypothesise a strong contemporaneous link between costs and reliability for TOCs operations. However, we do not discard the possibility of spill-overs across years: e.g. TOCs can invest today in future additional rolling stock capacity. Whilst outside the scope of this paper, cost-quality relationships for rail infrastructure would see spill-overs between time periods to a much greater extent.

5. Methodology and data

We now set out the estimation methodology and associated data, and link this back to the theoretical framework discussed in Sections 3 and 4.

We specify an econometric cost model that allows us to: i) observe empirically the relationship between cost and delay minutes for train operating companies and ii) estimate the marginal cost of delays reduction (i.e. MC of quality improvements; MCq).

The model assumes that all controllable costs (C_{it}) of decision-maker i (in our case, train operator) in period t (in our case, t refers to the year of operation) are explained as a function of outputs (y_{it}), a set of input prices (p_{it}) and the level of quality achieved (q_i). Our approach follows the literature (see for example, Kumbhakar et. al., 2015; Mizutani and Uranishi, 2013; Nash and Smith, 2007), with the innovation being in the treatment of quality within this framework in a railway application. Other explanatory variables of the model are: a time trend (t) and a set of variables accounting for heterogeneity among TOCs (z_{it}). Each of the elements of the cost function is described in detail below.

$$C_{it} = C(y_{it}, p_{it}, q_{it}, t, z_{it} \tag{2}$$

As discussed in Section 6 we adopt a logarithmic functional form for the cost function. C_{it} represents the total costs of train operating companies excluding any transfers to the Infrastructure Manager (in relation to access charges for access to the rail infrastructure, and performance penalties/compensation between TOCs and Network Rail). Thus compensation direct to passengers as a result of delays are included in TOC costs. On the other hand, Schedule 8 compensation payments are excluded because they reflect payments to / from Network Rail[5] (and indirectly, other operators) due to assumed long-term loss revenue as a result of delays caused by all operators and Network Rail – and these are therefore outside the production / cost function of the TOC under analysis. Jamasb et al. (2012) followed a similar approach when estimating MCq of UK electricity firms.

Quality (q_{it}) is represented by the minutes of delay caused by TOC i on itself, since these effectively represent the dimension of quality that TOCs are responsible for. Delays caused to other operators or on the TOC by other TOCs (or the infrastructure manager) are not relevant for our purposes.

A potential concern over endogeneity could appear to arise here to the extent that TOCs choose a desired level of quality. However,

[5] In some cases these are therefore "negative costs" to the TOC.

this is substantially mitigated in our case because, as noted in Section 4, the way that quality management is dealt with in respect of rail services in Britain – as with many other regulated industries – is through targets set by the regulatory system. In the case of rail operators, targets for delay minutes caused by operators are set out in the franchise agreements signed between the operator and the relevant national, or regional franchising authority. Thus to an extent operators can be seen to be aiming towards hitting a set of regulatory targets that are set exogenously. Jamasb et al. (2012) note the regulatory targets set for the energy sector, and the idea that quality is an attribute of great interest to policy makers (and thus regulated) in the railways can be seen as well in earlier railway papers such as Spady and Friedlander (1978).

Of course, ultimately there is no perfect exogenous variable in econometric estimation. In the case of quality, at the margin, train operators do face a performance regime which both penalises performance below target and which could give incentives to deliver performance above target. In this sense quality cannot be seen as entirely exogenous, although the regime does penalise companies from seeking to minimise cost by shading quality, which should reduce incentives to do so. It should also be noted that other variables typically included in rail cost functions, such as train-km, can also, in a similar way, be seen to partly exogenous (set by the franchise agreement) and partly at the discretion of operators seeking to expand services above the franchise specification. Such variables are routinely treated as exogenous in the literature. As noted above, the literature notes the role of regulation in setting exogenously given targets for quality measures. Other papers in this area (e.g. Gutacker et al., 2013) have noted the possibility of an endogeneity problem, though they argue that their (health) context also suggests the problem may not be large for similar reasons to those set out above.

However, as an additional safeguard, we use fixed effect estimation which explicitly recognises the possibility of correlation between operator specific effects (time invariant) and the included explanatory variables (e.g. delays). This approach therefore allows for the fact that (time invariant) omitted variables reflecting the different approaches and characteristics of different operators might be correlated with quality – whilst ensuring consistent estimates of the coefficients on the included explanatory variables (including delays); see, for example, Wooldridge, 2002[6]. This approach does not fully address the issue, since it only deals with the time invariant dimension of the potential endogeneity. Overall, therefore, we see the central problem at hand to be one in which TOCs are seeking to achieve a delay performance consistent with regulatory targets, and face the task of finding the appropriate balance between preventative and reactive costs to minimise overall costs. At the same time we recognise that operators may to some extent seek to choose their exact position relative to targets, and this is (at least partly – see above) dealt with by using fixed effects estimation to recognise this potential endogeneity.

Under the specification set out above, the elasticity of cost with respect to quality can be calculated as:

$$\varepsilon_q = \frac{\partial ln(C)}{\partial ln(q)} \tag{3}$$

The marginal cost of reducing minutes of delay caused on self for TOC i in period t can be calculated as follows (note that the negative sign is applied because delay minutes is an inverse measure of quality; see Jamasb et al., 2012):

$$MC_q = -\varepsilon_q \frac{C_{it}}{q_{it}} \tag{4}$$

where ε_q is the elasticity of the TOC-caused minutes of delay, equal to $\partial \ln C_t / \partial \ln q_t$ in our model, and q_{it} is the TOC-caused minutes of delay in year t by TOC i.

Several issues must be clarified to understand what can be learned empirically from the model. First, it is not possible to separate contemporaneous marginal proactive costs and marginal reactive costs of quality in estimation as separate data on the two aspects are not available. Only the overall marginal cost of quality (MCq) - i.e. of saving 1 min of delay - can be estimated (and not its proactive and reactive elements). Empirically, this is not a problem, since our objective is the estimation of the total marginal cost of quality. We expect to find $MC_q > 0$, which implies a negative coefficient for $\ln(q_{it})$ (reducing delays should be costly overall). However, $MC_q < 0$ would also be possible if firms were not efficient or if there were constraints in the industry structure. Therefore, a quadratic term for $\ln(q_{it})$ will be used to allow for both possibilities in line with the U-shape cost-quality curve from Fig. 3. Thus, understanding the two elements of cost (proactive and reactive) is important in understanding and explaining the observed relationship between cost and quality, even if the underlying cost elements remain unobserved.

Secondly, there may be non-contemporaneous effects, and the model can be adapted to capture these. Lags and leads of quality (e.g. q_{it-1}, q_{it+1}) can be added as explanatory variables. The quality from the previous period (q_{it-1}) can lead to spill over reactive effects. If anything, we might expect that the lower the delays last year, the lower the costs this year (positive coefficient for q_{it-1}). On the other hand, if firms proactively invest this year with the aim of reducing delays next year, there may be a significant negative relationship between C_{it} and q_{it+1} (negative coefficient for q_{it+1}).

Thirdly, relationship between costs and quality can vary by firm depending on their context and characteristics. For example, some TOCs may find it more expensive to reduce delays than others. Similarly, some TOCs may face more costly consequences if delays occur than others. This means the cost elasticity with respect to delay minutes can be allowed to vary by TOC characteristics to pick up heterogeneity. The elasticity may depend on: the salaries paid to staff, the load of the trains (more passengers mean more compensation saved if delays are reduced), the density of their operation, the length of their trains (longer trains require more resources for inspection and maintenance) or the number of stations operated. The model can therefore accommodate context-dependent elasticities

[6] See page 638.

and marginal costs via interaction terms between TOC characteristics and q_{it}.

The specific variables (other than delays) included in the cost function are standard in the rail literature (see for example, Mizutani and Uranishi, 2013; Wheat and Smith, 2015). To summarise, the set of TOC outputs is formed by: train density (train-km per route-km), average length of train (vehicle-km per train-km) and number of stations operated. Route-km is also included in the estimation to distinguish between scale and density effects (Smith and Wheat, 2012). The cost function estimated is standard for the rail literature as noted. The set of input prices contains the average salary in the firm and an input price representing other costs (expressed as an average price per rolling stock numbers). Other variables (z_{it}) included in the dataset are load factor per vehicle (passenger-km/vehicle-km), the type of service dummies (intercity, regional and LSE) and the delay minutes caused by TOCs which is our reliability metric. Descriptive statistics are in Table 1 below.

The final sample excluded one TOC (London Overground) out of the 18 available, after inconsistencies were detected in the process of data validation contrasting various sources. The final sample contains a total of 85 observations across the remaining 17 TOCs over the 5-year period. The sample size is in line with the range of relevant published work in the academic and regulatory literatures (see for example, Jamasb et al., 2012; CEPA, 2014) – economic regulators often have to work with relatively limited sample sizes. As shown below, the sample size does not hinder the estimation of statistically significant and plausible coefficients.

6. Results and discussion

We follow the literature by starting with the widely used translog cost function (see for example, Mizutani and Uranishi, 2013; Kumbhakar et al., 2015). Our preferred cost model is a restricted version of the translog functional form (see e.g. Smith and Wheat, 2012); retaining only a subset of the squared and interaction terms, as their inclusion was not supported by the data. Starting from a full translog, multiple search specification paths were tested to confirm the results were robust. This restricted translog is also preferred to a simpler Cobb-Douglas model as indicated by a likelihood ratio test. We note that the key findings in respect the relationship between cost and quality is not sensitive to our choice of functional form.

The model estimates are shown in Table 2. All variables are specified in logs after being scaled by the sample mean as is standard in the literature[7]. Consequently, all first-order coefficients can be interpreted as elasticities at the sample mean. It should also be noted that homogeneity in input prices is imposed in the model, by using one input price to scale the dependent variable and the second remaining input price (salary).

As discussed in Section 5, the preferred model uses a fixed effects (FE) approach. Fixed effects permits the estimation of consistent parameter estimates even in the presence of correlation between the regressors and the firm effects (which could occur to the extent that quality is at least partly endogenous). Further, we would like to isolate the within-firm temporal variation in cost in relation to quality (see e.g. Jamasb et al., 2012). This FE choice is also supported empirically: the Hausman test rejected the Random Effects (RE) specification in favour of FE. Importantly, the outcomes in terms of the quality variable findings are similar under both approaches, so the choice of FE or RE is not determinant. The panel approach used is standard in the econometric cost modelling literature in general and also in rail and other transport applications (see for example, Kumbhakar et al., 2015; Nash and Smith, 2007).

All output and input price coefficients are estimated with the expected sign. The majority of coefficients are statistically significant. Although the coefficient on the stations variable is not statistically significant, we rejected the null hypothesis that all stations-related coefficients (including interactions) is zero via a Wald test. The results are within the expected ranges from the past literature (e.g. Smith, 2006; Smith and Wheat, 2012; Wheat and Smith, 2015). The models show the expected high levels of explanatory power in line with the past literature, bearing in mind that we estimate a logarithmic cost function. We tested some lagged relationships in the model which were not found to be statistically significant; in particular we could not find evidence of a lagged relationship between costs and delays. This finding is broadly in line with our expectation that, because we are focusing on train operations (not infrastructure) the relationship between cost and delay is more likely to be within-year (contemporaneous); see Section 4.

Moving on to the quality findings, the first-order coefficient on delay minutes (DMT) is negative and significant at the 99% level of confidence. This means that, at the sample mean, the cost elasticity of delay minutes is negative, i.e. reducing delays is associated with higher costs. The full relationship between cost and delays across the whole sample is captured by this and the interaction parameters. To interpret the results with reference to the theoretical discussion (i.e. in terms of cost elasticity of quality), we need to reverse the sign of the elasticity of delay minutes (delay is an inverse measure of quality). Thus, the cost elasticity of quality is positive at the sample mean, which means it is costly to improve quality. In other words, the average firm faces a positive marginal cost of quality ($MCq > 0$).

With respect to the squared and interaction terms, the squared term takes the reverse sign to the first-order coefficient, implying a quadratic relationship between cost and quality. That is, since the elasticity of cost with respect to quality depends on the first order coefficient on quality, as well as on the second order terms, the reverse sign on the squared term means that the cost elasticity of quality (and in turn MCq) – which is positive at the sample mean – becomes negative as quality falls. There is also heterogeneity in the elasticity across TOCs, which varies with train length, salary, load per vehicle and number of stations. The direction of these effects is plausible as discussed below. Taking together all coefficients on delays to estimate the elasticity (ε_q) and the MCq for the sample, the results match the expected theoretical shape of the cost-quality relationship (see Figs. 5 and 6).

We calculate the cost elasticity of quality (ε_q) for each TOC and plot it against delay minutes per train planned (Fig. 5). This relative measure is a more representative quality scale for TOCs. Note that the horizontal axis (delay minutes per train planned) has been

[7] For the zero values of the stations variable we take the logarithm of a small value.

Table 1
Descriptive statistics.

Variable	Description	Mean	Sd.Dev.	Min	Max
EXP	Total annual controllable expenditure, i.e. excludes track-access charges; in million £ (£,constant prices)	271.7	155.6	58.7	755.5
SAL	Average salary (k£)	33.1	3.4	26.3	45.6
OTHER	Average price of other input (k£)	29.9	14.0	12.9	61.1
TRAINKM	Train-km per year (million)	29.9	14.8	6.1	63.8
PASSKM	Passenger-km per year (million)	3375.9	2143.8	545.1	8691.0
ROUTEKM	Route-km	1355.1	854.3	115.5	3065.8
DENS	Density (train-km per year/route-km) (million)	0.029	0.016	0.012	0.056
STATIONS[a]	Number of stations operated	143.6	127.2	0.0	464
LOAD	Average Passengers per train per year (PASSKM/TRAINKM)	114.8	45.6	45.6	240.6
AVLEN	Average length of train (Vehicle-km/TRAINKM)	5.3	2.2	2.5	11.5
LOADVE	Average Passengers per vehicle per year (LOAD/AVLEN)	22.2	4.0	16.4	33.2
DMT	Annual Delay minutes caused by TOC	199,461	162,820	14,065	672,567
INTERCIT	Intercity dummy[b]. Omitted dummy: regional services	0.4	0.4	0.0	1.0
LSE	London and South East dummy.	0.3	0.5	0.0	1.0

[a]Note that one of the operators does not run any stations (i.e. they are either run by the infrastructure manager or another operator).
[b]These standard categories are used in reporting rail data in Great Britain, reflecting the different types of services: Long distance (intercity), services in London and the South East, and regional services.

Table 2
Preferred model results.

Variable†	Est.	Std.E		t-ratio
SAL	0.497	0.027	***	18.18
ROUTEKM	0.578	0.178	***	3.24
DENS	0.774	0.168	***	4.62
STATIONS	0.311	0.309		1.01
LOADVE	0.210	0.119	*	1.77
AVLEN	0.358	0.147	**	2.43
T	−0.101	0.046	**	−2.19
T^2	0.004	0.002	**	2.26
LOADVE^2	0.698	0.191	***	3.66
DMT	−0.077	0.029	***	−2.69
DMT^2	0.036	0.013	***	2.66
DMT*STA	−0.030	0.012	**	−2.54
DMT*SAL	0.053	0.021	**	2.53
DMT*LEN	−0.135	0.051	**	−2.64
DMT*LVE	0.259	0.077	***	3.37
Constant				
R-squared	0.99			
Observations	85			

*, **, *** indicates significance at the 90%, 95% and 99% confidence level respectively. † All variable names correspond to those described in Table 1; T is the time trend and T^2 is the squared value of the timetrend. The interaction terms with DMT refer to interactions with the variables: STATIONS, SAL, AVLEN and LOADVE respectively.

reversed in order to reflect quality from left to right as in the theoretical discussion (i.e. zero delays is at the right-hand side of the graph). This allows us to compare the empirical results with the theoretical framework. Note we distinguish two of the TOCs with red dots (East Coast and West Coast inter-city franchises) to assist the explanation of their differentiated results.

The cost elasticity with respect of quality estimates are within the range −0.1 to 0.15 for most TOCs, and around 0.2–0.25 for the East and West Coast TOC franchises (see Fig. 5). In general, the elasticity increases with quality. The negative estimates for some TOCs further indicate that in some cases it is feasible to improve quality and reduce costs simultaneously, in line with the theoretical discussion earlier. This can be the case of TOCs with a sub-optimal strategy that leads to high delays and high reactive costs. The results also suggest a different cost-quality relationship for the East and West Coast TOC franchises, which happen to be the fastest intercity services running on two core corridors of the network. The possibility of cost-quality spill-overs over time was tested using leads and lags on the delay minutes variable, but none of these terms were found to be significant; hence the cost-quality relationship is highly contemporary in the context of train operating companies as expected.

Marginal cost of quality (MCq) estimates

The marginal cost of quality for each TOC i in period t can be obtained using the calculated elasticities of quality (Eq. (4)). The estimates are reported in Fig. 6 and plotted also against the number of delay minutes per trains planned (the right-hand side of the horizontal axis corresponds to highest quality, i.e. zero delays). While the MCq is initially calculated per train, we divide the MCq by the average train load to obtain an estimate of the MCq per passenger.

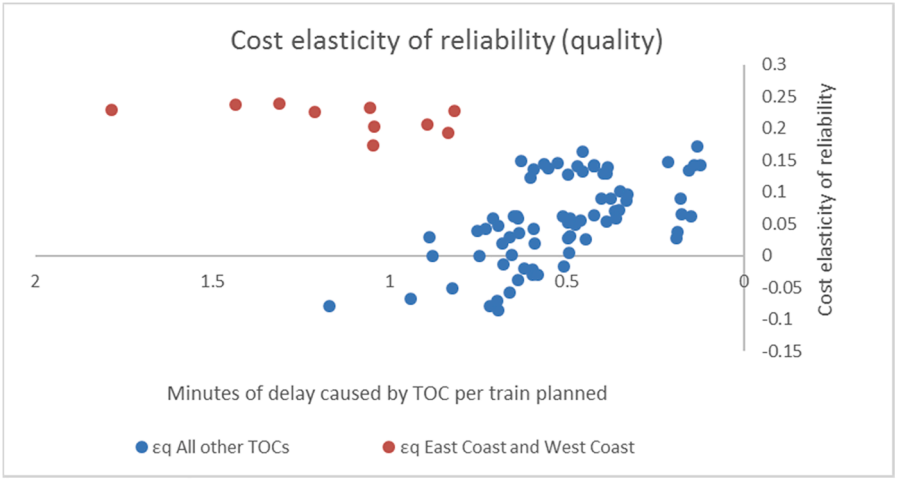

Fig. 5. Estimates of cost elasticity of quality (quality = reliability).

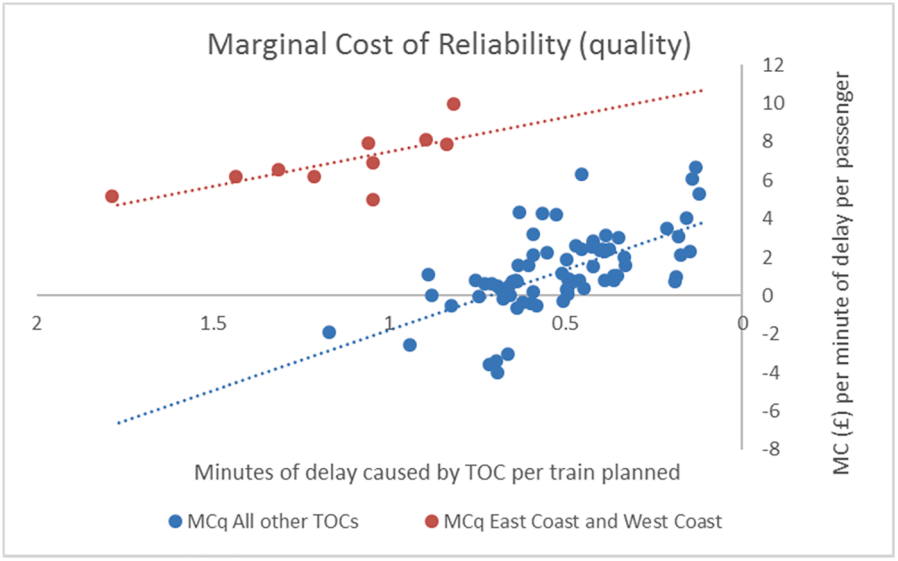

Fig. 6. Estimates of marginal costs of quality.

Overall, the results conform to the theoretical expectations: the estimated MCq curve is upward sloping with respect to quality. Most estimates fall within the right-hand side of the graph (recalling Fig. 3), where MCq > 0. This means that most companies are, at least, producing a quality level greater that Q_{min}, as expected. However, in roughly 18% of the cases, quality is lower than Q_{min} meaning that TOCs could improve quality and reduce costs simultaneously. In these cases, TOCs are making reliability decisions sub-optimally.

The MCq is, on average, just below 2 £/min per passenger. The distribution of MCq ranges from −4 £/min to 10 £/min, with the majority of estimates ranging from −0.5 to 6 £/min. We note that there are no similar estimates from the literature against which to compare these findings as they are the first in the literature. Future research could usefully seek to compare the results from the model with bottom-up analysis of scheme-by scheme investments designed to improve quality on the railways.

There are several possible explanations as to why a TOC may encounter a negative MCq. As discussed earlier, the MCq estimates hide a mix of proactive and reactive marginal costs that we cannot disentangle. A TOC with negative MCq is facing the possibility to invest in quality and make a profit on the investment via savings on reactive costs. Hence, one naive explanation is that a TOC in this position has not realised this possibility. In other industries, some researchers have argued that it is likely to find firms in this situation (see e.g. Schiffauerova and Thomson, 2006). For example, a TOC might "gamble", trying to spend little to achieve high quality; it is only after failing they find themselves expending more, in a reactive manner (having then delivered poor quality).

On the other hand, another hypothesis for MCq < 0 is the presence of constraints in a vertically separated industry. This would mean that, although a TOC realises the potential for cost savings and quality improvement, it is limited in terms of the actions that

would unlock greater reactive cost savings. These aspects could form the basis for future research.

The observed heterogeneity in the estimates implies that different TOCs may face a different cost-quality relationships:

i) MCq increases with train length. This result may reflect that preventive inspection and maintenance costs are higher for trains comprising more vehicles.

ii) MCq is lower if salaries are higher. Staff costs can be both proactive (e.g. staff related investments) and reactive (e.g. staff extra hours). This result shows that the reactive component has relatively more weight in our data, and hence improving quality is 'cheaper' when salaries are high thanks to the greater potential for reactive salary cost savings.

iii) MCq is lower if load per vehicle is higher. This results reflects that reducing delays turns out to be cheaper when vehicles are fuller, probably due to the greater savings in passengers' compensations (reactive costs).

iv) MCq is higher when the number of stations managed is higher. The link between stations and MCq is not as intuitive as the previous ones. We do not have a clear rationale for this finding; we also note that our functional form is trying to approximate a potentially complex technology.

As noted above, there is no other evidence in the literature that is comparable to these estimates. Therefore, as an illustration to put our MCq estimates into context, we plot them jointly with estimates of the marginal benefit of improving quality (Fig. 7). We calculate the marginal benefits as the willingness-to-pay (WTPq) for quality improvement, i.e. for reducing delays for rail passengers, using the results from the most recent national study on time and reliability valuation in the UK (ARUP et al., 2015). These WTPq per delay minute are available per passenger, and therefore are comparable with our marginal cost per passenger estimates[8].

The comparison shows that, on average, the MCq is slightly higher than the WTPq. This indicates that in most cases, at the margin, it is more expensive to reduce delays than the benefits that come through valuations based on WTP. Thus, a straightforward comparison of the estimated marginal costs from our model against indicative measures of willingness to pay for such improvements might suggest that in many cases delay improvements are not worthwhile. That said, there are several caveats to such interpretation. First, the WTP metrics do not represent all the benefits. Second, we see that there is wide variability in the estimated MCq – much more so than for typical WTP values. This is expected given that the method adopted in our paper imply that marginal costs are likely to vary by TOC, whereas the WTP estimates are averages for three different TOC types (intercity; London and South East and regional services; and mixed TOCs). Third, in practice, specific interventions that reduce delays in particular circumstances could well produce different marginal cost estimates to those from our econometric top-down model, and it is the former that would ultimately be the relevant costs for any economic appraisal; which would also need to consider costs and benefits to other TOCs and also to the infrastructure manager, Network Rail. WTP measures could also then vary depending on the specific flows impacted by the interventions.

However, what our results show is that any economic appraisal of interventions designed to improve performance should:

i) Take into account that some TOCs may see significant (reactive) cost savings from improved delay performance, which could easily be overlooked otherwise. Such situations will quickly see benefits from improving quality, even before passenger benefits are considered.

ii) Consequently, consider assessments of quality improvement interventions on a case-by-case basis, to reflect the very different impacts for TOCs at different ends of the scale (very poor vs. very high quality). In this respect, our results suggest that poorly performing TOCs could even see improved performance alongside overall cost reductions.

The main take-away of the results is that they present initial evidence on the hypothesised non-linear relationship between costs and delays: often high quality and high cost go hand in hand, but in some instances we observe lower quality that is linked to higher costs.

Implications for performance incentives

This results in this paper can also contribute to the regulatory discussion around the design of performance incentives in vertically separated industries, both in respect of the theoretical framework and the empirical results. The proposed framework highlights that the marginal cost of quality is determined not only by how costly it is to implement delay-reduction actions, but also by how much reactive costs might be saved. This means that we can envisage two types of quality incentives: 'natural' and 'imposed'. Imposed incentives are those established by the regulator to ensure a certain level of punctuality. Natural incentives are those naturally faced by the firms regardless of the regulatory environment: any TOC should at least provide quality at Qmin level (see Fig. 3), as otherwise it would face a lose-lose situation with higher costs and poor quality.

The design of performance incentives should consider these natural incentives. Herein, our results show that it is possible to estimate the level of Qmin, which may be different across TOCs. Thus, imposed incentives might only be justifiable above and beyond Qmin, as otherwise there is a risk of subsidising an inefficient firm (e.g. Gutacker et al., 2013). Water regulator Ofwat has also faced similar issues – namely that high cost is associated with poor performance, and thus the inclusion of quality variables in a cost function used for benchmarking purposes would award a higher cost allowance for poor performance (which would be considered perverse);

[8] WTP for delay savings are, in essence, obtained by asking people how much they would be willing to pay to receive more punctual services (ARUP et al., 2015). The WTP presented here are estimated by multiplying values of travel time savings (VTT) by the so-called lateness multiplier, often found to be around 3 (this is, one minute of delay is equivalent to 3 scheduled minutes; see Wardman and Batley, 2014). For the VTT values, we use the rail specific values by trip purpose (ARUP et al., 2015), combined with national aggregate purpose shares. For the TOCs covering Regional and London and South East (LSE) services, this gives that one minute of delay (per passenger) should be valued at 0.62 £/min. For Intercity services, we use a higher purpose share for business trips (associated with much larger VTT), giving a value of 1.54 £/min (one TOC has a mix of services and therefore has a different valuation). While WTP varies with the assumptions above, these values reflect a typical range for WTPs for delay savings (e.g. see Wardman and Batley, 2014) and it serves the illustration purpose.

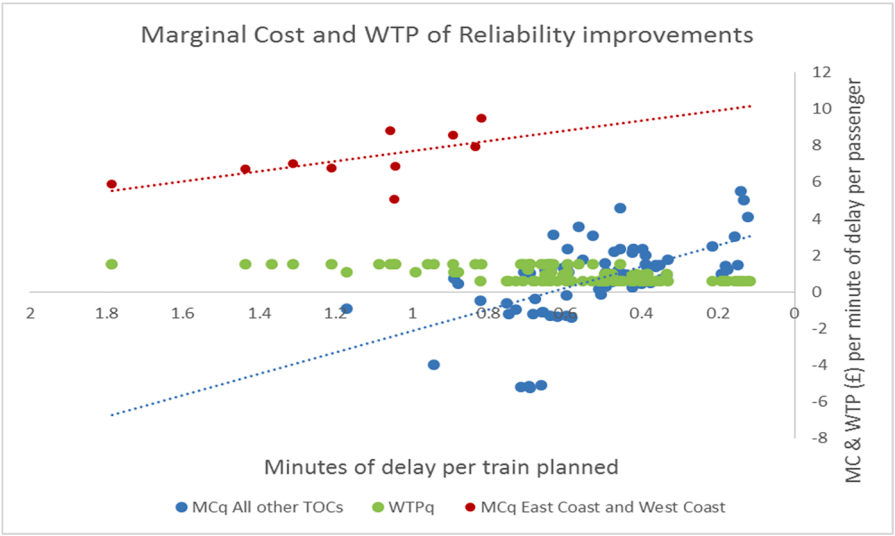

Fig. 7. Marginal costs and WTP for quality improvements.

see CEPA (2014).

Another implication of the framework is that reactive costs, and hence incentives, depend on the consequences faced by TOCs. One interesting aspect of this is the compensation to passengers for the delays suffered. The passenger compensation scheme means that TOCs are obliged to compensate passengers in Great Britain. There is a threshold of delay minutes (30 min)[9] below which passengers are not entitled to compensation (Office of Rail Regulation, 2013). The framework shows that higher reactive costs act as an incentive to improve quality. Therefore, the regulator could induce higher levels of quality simply by ensuring these reactive costs are incurred. However, the reality shows that only 1 out 3 passengers claim their entitled compensation (Office of Rail Regulation, 2013). If current compensation channels are too onerous for passengers (TOCs do not have a short-term incentive to facilitate this), facilitating compensation would promote a naturally higher level of Qmin (see Fig. 8, movement from Qmin_0 to Qmin_1). Indicating a step in this direction, in Great Britain, passenger compensation has become progressively easier to claim and in some case is automated for tickets purchased on operator travel applications.

Fig. 8 depicts the effect of facilitating compensation payments to passengers, which basically imposes higher reactive costs on TOCs (MRCq shifts downwards). Under such scenario, when a TOC contemplates a quality improvement they would also observe a higher reactive cost saving (the MRCq curve is more negative for any level of quality). Hence, TOCs would therefore observe, for a given level of quality, a lower MCq for further improvements in quality (MCq curve shifts downwards as MCq = MPCq + MRCq). So Qmin moves to the right; and the marginal cost of improving quality is also lower than before, for improvements in quality above that level.

7. Conclusions

This paper has explored the relationship between travel time reliability and the costs of train operating companies. Our work makes a number of important contributions.

1. The paper has brought together several bodies of literature which had remained disconnected. We have connected a central theme of ‘costs and quality’ that applies across literatures on railway and transport economics, energy economics, health economics and management. Railway research has paid very little attention to the relationship between costs and quality; Cost of Quality models have been developed in the management literature but not from a marginal costs perspective; and a few studies in energy and health have studied the marginal costs of quality, but these were context-specific and remained unaware and disconnected from existing Cost of Quality models.
2. Secondly, building on the past literature, we set out a framework for the cost-quality relationship from a marginal costs perspective so that it can be directly applied for estimation purposes and to aid firms’ understanding and decision-making and regulatory practices (e.g. quality incentives design). Our framework proposed that marginal cost of quality is a combination of marginal proactive costs and marginal reactive costs. Our focus is on the understanding of this framework in the context of railways, but the framework is general enough to be used in other contexts, in and outside transportation.

[9] The 30 min threshold applied during the dataset employed in this paper. From 2016, changes were made to reduce the threshold to 15 min in some cases.

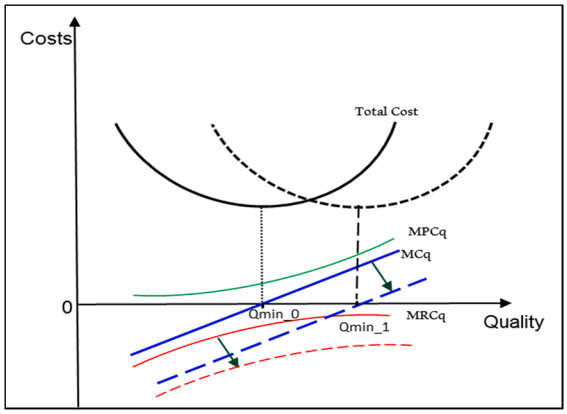

Fig. 8. Effect of reinforcing delay consequences.

3. Third, we provide the first estimates on the cost elasticity and marginal cost of reducing delays for train operating companies in the literature. The empirical work has shown that in most cases TOCs face a positive marginal cost of quality which, assuming that firms are operating optimally, is expected. At the sample mean, we found that it is costly to reduce delays for TOCs. The cost elasticity of quality was on average around 0.08, but it varied with the level of quality, within and across TOCs. TOCS characteristics such as train length, vehicle load or salaries were found to influence the estimated elasticities. Overall, the marginal cost of quality increased with the level of quality, as theoretically expected.
4. Finally, our analysis also revealed that the marginal cost of quality was negative in approximately 18% of the cases. In those cases, reducing delays would also bring cost savings, and the obvious follow-up question would be to understand why TOCs would not exploit such opportunity. That is, why are some TOCs operating on a sub-optimal part of the cost-quality curve? One explanation might be the presence of industry structure constraints that prevent TOCs from doing so. Linking back to the title of the paper: achieving higher quality is not always costly. Changes to incentive regimes could encourage improved quality performance.

This paper is one step to improve our understanding of quality in railway supply. Both the theoretical framework and the empirical application can be used to aid better decision-making in the railway and assist the design of incentives systems in relation to quality aspects. For example, ensuring increased take-up by passengers of passenger compensation schemes (which can often be low) would increase the marginal reactive cost of delays and thus act as an incentive for operators to improve performance. Travel time reliability is one aspect of quality, but the framework can easily be translated to other quality contexts. Bottom-up approaches working in-depth with TOCs would be highly welcome to further explain, contrast and complement the first estimates of marginal cost of delay reductions provided in this paper; alongside further top-down econometric work. Finally, we also suggest further research on approaches to address the potential endogeneity issue that can arise when including quality measures in cost functions.

Credit authorship contribution statement

Andrew S.J. Smith: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Writing – original draft, Writing – review & editing. **Manuel Ojeda Cabral:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This paper was written as part of the EU Horizon 2020 funded NeTIRail-INFRA project. The authors are also thankful to Phill Wheat, Alex Stead (University of Leeds, UK) and Jan-Eric Nilsson (VTI, Sweden) for their useful comments on this work.

References

Abate, M., Lijesen, M., Pels, E., Roelevelt, A., 2013. The impact of reliability on the productivity of railroad companies. Transp. Res. Part E 51, 41–49.
Affuso, L., Alvaro, A. and Pollitt, M., 2002. Measuring the Efficiency of Britain's Privatised Train Operating Companies. No. 48: Regulation Initiative Discussion Paper Series.
Assaf, A.G., Josiassen, A., Gillen, D., 2014. Measuring firm performance: Bayesian estimates with good and bad outputs. J. Business Res. 67, 1249–1256.

ARUP, Institute for Transport Studies University of Leeds and Accent, 2015. Provision of Market Research for Value of Time Savings and Reliability. Phase 2 Report to the Department for Transport, https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/470231/vtts-phase-2-report-issue-august-2015.pdf, accessed 10/05/19.
Australian Competition and Consumer Commission, 2012. Regulatory Practices in Other Countries: Benchmarking opex and capex in energy networks.
Carrion, C., Levinson, D., 2012. Value of travel time reliability: A review of current evidence. Transp. Res. Part A-Policy Practice 46 (4), 720–741.
Cascetta, E., Cartenì, A., 2014. A quality-based approach to public transportation planning: theory and a case study. Int. J. Sustain. Transp. 8 (1), 84–106.
CEPA, 2014. Ofwat Cost Assessment – Advanced Econometric Models.
Coelli, T.J., Gautier, A., Perelman, S., Saplacan-Pop, R., 2013. Estimating the cost of improving quality in electricity distribution: A parametric distance function approach. Energy Policy. 53, 287–297.
Crosby, P.B., 1979. Quality is Free. New York: McGraw-Hill.
dell'Olio, L., Ibeas, A., Cecin, P., 2011. The quality of service desired by public transport users. Transp. Policy 18 (1), 217–227.
Estache, A., Perelman, S., Trujillo, L., 2007. Measuring quantity-quality trade-offs in regulation: the Brazilian freight railways case. Ann. Public and Cooperative Econ. 78 (1), 1–20.
Feigenbaum, A.V., 1956. Total quality-control. Harvard Business Rev. 34 (6), 93–101.
Freeman, J.M., 1995. Estimating quality costs. J. Oper. Res. Soc. 46 (6), 675–686.
Kennedy, J., Smith, A.S.J., 2004. Assessing the efficient cost of sustaining Britain's rail network. Perspectives based on zonal comparisons. J. Transport Econ. Policy 38 Part 2, 157–190.
Gutacker, N., Bojke, C., Daidone, S., Devlin, N.J., Parkin, D., Street, A., 2013. Truly inefficient or providing better quality of care? Analysing the relationship between risk-adjusted hospital costs and patients' health outcomes. Health Econ. 22 (8), 931–947.
Hart, Shleifer, Vishny, 1997. The proper scope of governance. Quarterly J. Econ., 112 (4): 1127-1162.
House of Commons Transport Committee, 2017. Rail Franchising: ninth report of session 2016-17. Available at https://www.publications.parliament.uk/pa/cm201617/cmselect/cmtrans/66/66.pdf, last accessed 10/05/2019.
Hwang, G., Aspinwall, E., 1996. Quality cost models and their application: a review. Total Qual. Manag. 7 (3), 267–282.
Jamasb, T., Orea, L., Pollitt, M., 2012. Estimating the marginal cost of quality improvements: The case of the UK electricity distribution companies. Energy Econ. 34 (5), 1498–1506.
Juran, J.M., 1951. Juran's Quality Handbook, 1st edn (New York: McGraw-Hill).
Juran, J.M., Gryna, F.M., 1988. Quality Control Handbook, 4th Edn, New York: McGraw-Hill.
Juran, J.M., Seder, L.A., Gryna, F.M., 1962. Quality control handbook, 2nd ed. McGraw-Hill Book Company, New York, NY.
Kumbhakar, S.C., Wang, H.-J., Horncastle, A.P., 2015. A Practitioner's Guide to Stochastic Frontier Analysis Using Stata.
Lawrence, D., Coelli, T., Kain, J., 2014. Economic Benchmarking Assessment of Operating Expenditure for NSW and ACT Electricity DNSPs.
Link, H., 2019. The impact of including service quality into efficiency analysis: The case of franchising regional rail passenger serves in Germany. Transp. Res. Part A: Policy Practice 119, 284–300.
Masser, W.J., 1957. The quality manager and quality costs, Industrial Quality Control, October, pp. 5-8.
Mayat, T., Wheat, P.E., 2020. Do the Public Perceive Good Asset Management? The Case of Local Highways in England, Mimeo, University of Leeds.
Merkert, R., Assaf, A.G., 2015. Using DEA models to jointly estimate service quality perception and profitability – Evidence from international airports. Transp. Res. Part A 75, 42–50.
Mizutani, F., Uranishi, S., 2013. Does vertical separation reduce cost? An empirical analysis of the rail industry in European and east Asian OECD countries. J. Regul. Econ. 43 (1), 31–59.
Nash, C.A., Smith, A.S.J., 2007. Modelling Performance: Rail. In: Hensher, D.A., Button, K.J. (Eds.), Handbook of Transport Modelling. Elsevier, Second Edition.
Network Rail (2016). Website http://www.networkrail.co.uk/timetables-and-travel/delays-explained/.
NEXTOR, 2010. Total Delay Impact Study. A comprehensive assessment of the costs and impacts of flight delay in the United States.
Ofgem, 2021. RIIO-2 Final Determinations - Core Document.
Ofwat, 2019. PR19 final determinations: Overview of final determinations.
Office of Rail Regulation, 2013. Rail passenger compensation and refund rights. Available at http://www.orr.gov.uk/_data/assets/pdf_file/0003/10668/passenger-compensation-refund-rights-aug-2013.pdf, last accessed 10/05/2019.
Peimbert-Garcia, R.E., Limon-Robles, J., Beruvides, M.G., 2016. Cost of quality modeling for maintenance employing opportunity and infant mortality costs: An analysis of an electric utility. Eng. Econ. 61 (2), 112–127.
Peterson, E.B., Neels, K., Barczi, N., Graham, T., 2013. The economic cost of airline flight delay. J. Transport Econ. Policy. 47, 107–121.
Rail Delivery Group, 2018. Country Profiles – Japan: The Williams Review, London. https://www.raildeliverygroup.com/files/Publications/consultations/williams/2018-12-21_rdg_submission_williams_japan_profile.pdf.
Schiffauerova, A., Thomson, V., 2006. A review of research on cost of quality models and best practices. Int. J. Quality Reliability Manage. 23 (6), 647–669.
Shift2Rail, 2015, Shift2Rail Multi-Annual Action Plan: Executive View Part A.
Smith, A.S.J., 2006. Are Britain's railways costing too much? Perspectives based on TFP comparisons with British Rail 1963–2002. J. Transport Econ. Policy 40, 1–44.
Smith, A.S.J., Wheat, P., 2012. Evaluating alternative policy responses to franchise failure evidence from the passenger ran sector in Britain. J. Transport Econ. Policy 46, 25–49.
Spady, R.H., Friedlander, A.F., 1978. Hedonic cost functions for the regulated trucking industry. Bell J. Econ. (Spring) 159–179.
Srivastava, S.K., 2008. Towards estimating Cost of Quality in supply chains. Total Quality Manage. Business Excellence 19 (3), 193–208.
Stradling, S.G., Anable, J., Carreno, M., 2007. Performance, importance and user disgruntlement: A six-step method for measuring satisfaction with travel modes. Transp. Res. Part A 41 (1), 98–106.
Sumicsid, 2007. Project Gerner / AS6: Development of Benchmarking Models for German Electricity and Gas Distribution: Final Report.
The Williams Rail Review, 2019. The user experience of the railway in Great Britain: Evidence paper, March 2019.
Transport Focus, 2020. National Rail Passenger Survey: Main Report Spring 2020.
TSLG, 2012. The Future Railway: The Industry's Rail Technical Strategy 2012, Supporting Railway Business.
van Oort, N., 2016. Incorporating enhanced service reliability of public transport in cost-benefit analyses. Public Transport 8 (1), 143–160.
Wardman, M., Batley, R., 2014. Travel time reliability: a review of late time valuations, elasticities and demand impacts in the passenger rail market in Great Britain. Transportation 41 (5), 1041–1069.
Wardman, M., Batley, R., 2021. The demand impacts of train punctuality in Great Britain: systematic review, meta-analysis and some new econometric insights. Transportation 1–35.
Wheat, P.E., Smith, A.S.J., 2015. Do the usual results of railway returns to scale and density hold in the case of heterogeneity in outputs: A hedonic cost function approach. J. Transport Econ. Policy. 49 (1), 35–47.
Wooldridge, J.M., 2002. Econometric Analysis of Cross-section and Panel Data, MIT Press, London.
Yu, W., Jamasb, T., Pollitt, M., 2009. Willingness-to-pay for quality of service: an application to efficiency analysis of the UK electricity distribution utilities. The Energy J. 30 (4) https://doi.org/10.5547/ISSN0195-6574-EJ10.5547/ISSN0195-6574-EJ-Vol30-No410.5547/ISSN0195-6574-EJ-Vol30-No4-1.