NeTIRail

Needs Tailored Interoperable Railway Infrastructure

Deliverable D1.7

Incentives final report

Submission date : 21 December 2017

NeTIRail-INFRA is funded by the European Commission under the Horizon 2020 programme - Grant agreement n° 636237

Lead contractor

University of Leeds

Project Coordinator

University of Sheffield, USFD

Executive Summary

This deliverable draws together a significant body of research undertaken as part of the project.

A quote below is taken from the description of works (DOW):

> *"…However, there is also a need to identify incentive mechanisms that will help deliver these benefits. This can be achieved through ensuring that different players operating in the rail industry have the right incentives to implement the proposed technical solutions. There are a range of incentive mechanisms operating within European rail systems. Of particular importance are the charging mechanisms for use of infrastructure (track access charging regimes), as these affect how much infrastructure managers are paid for usage of the infrastructure. The wider regulatory, organisational, governance and funding arrangements in which infrastructure managers (and other players) are operating is also important.*
>
> *The research frontier is to develop new understanding of the impact of climate and quality (e.g. reliability) on marginal and average costs; develop our understanding of cost variability and relative efficiency by exploiting new datasets that are more closely aligned to business units; and to extend the research on the impact of track access regimes on behaviour to incorporate the wider regulatory frameworks in which railways operate and how these impact on costs and efficiency. In developing all of these research strands best practice approaches from SUSTRAIL will be built into the approach. The proposed incentives research provides the link between technical solutions, implementation and impact."*

The research under this task falls under two broad headings with two broad objectives: (1) econometric modelling aimed at pushing forward the research frontier in the area of rail marginal (and average) cost estimation; and (2) qualitative research will be carried out in the area of incentives: that is how different incentive mechanisms (track access agreements, franchise agreements, performance regimes and the wider regulatory and government funding regimes) operate and interact / contradict.

These objectives have been met and the results are contained in five research studies are reported in this deliverable in the following areas:

a. Research on incentives for innovation (Deliverable 1.7 Annex 1)
b. The impact of quality on costs (Deliverable 1.7 Annex 2)
c. Methodological aspects of marginal cost modelling: Estimating the marginal cost of different vehicle types on rail infrastructure (Deliverable 1.7 Annex 3)
d. Methodological aspects of marginal cost modelling: Bayesian techniques (Deliverable 1.7 Annex 4)
e. Methodological aspects of marginal cost modelling: Dynamic techniques (Deliverable 1.7 Annex 5)

Annexes 2 to 4 are concerned with the first objective (advancing the research frontier in marginal (and average) cost estimation). It was noted in Deliverable D1.3 that in delivering the objectives in this area of work, prioritisation decisions would be made to place greater emphasis on different aspects depending on data availability and on where most progress could realistically be made. Thus the research focused on modelling the relationship between cost and quality and on three methodological areas, covering a combined engineering / econometric approach to estimating the marginal cost of different vehicle types using disaggregate cost data (Annex 3), Bayesian techniques to support the use of prior information in informing new studies of rail infrastructure cost variability (Annex 4) and dynamic models that recognise the time lags in the relationship between rail infrastructure cost and traffic and the inter-relationships between maintenance and renewals. Annex 1 covers the second objective (qualitative research in the area of incentives).

The policy motivation for developing the research frontier in these areas is clear. The advantages and disadvantages of alternative railway structures have long been the source of considerable debate. EU legislation requires a degree of separation of infrastructure from operations; however, some countries have adopted full, legal separation (e.g. as in Sweden and the UK), whilst others have adopted a holding company structure (most notably in Germany and more recently in France). Different structures may be expected to have differential impacts on rail industry costs and specifically with respect to the research in this project on innovation. Other factors in addition to structure (e.g. economic regulation; the role of government funding and multi-annual agreements; the terms and length of rail franchises) will also have an impact. Better understanding of the barriers to and enablers of innovation in complex rail organisational structures is therefore of great policy importance and these issues are explored through several case studies drawn from the countries covered by the NeTIRail-INFRA consortium (see Annex 1).

In respect of the cost of quality railways across Europe face the challenge of how to enhance performance whilst also improving productivity and reducing costs. Therefore, understanding more about the relationship between cost and quality (in our case as measured by delay minutes) is crucial: that is, how much does it cost to reduce delay minutes. This information can be combined with measures of willingness to pay, to determine an optimal level of quality performance which can then be reflected in performance regimes and other regulatory target mechanisms (see Annex 2).

Understanding the marginal cost of railway usage and different types of vehicle is also an important area of research from a policy perspective, since this information is needed to set track access charges for rail operators accessing the rail network. Accurate information on marginal cost should incentivise optimal use of the existing rail network – that is, traffic should not run if it cannot cover its marginal wear and tear cost – and differentiation of charges by vehicle type should incentivise the adoption of track friendly vehicles. Infrastructure managers in some countries use econometric methods to set access charges (e.g. in France), whilst others utilise engineering methods. The research reported in this deliverable advances both econometric techniques (annexes 4 and 5) and an approach that seeks to combine econometric and engineering approaches (annex 3).

This report contains an executive summary on the key results from each of the above research areas. Some concluding remarks are also offered to highlight that the results have both generalizable findings but also specific implications for the case studies considered in the NeTIRail-INFRA project. The implications of this deliverable to the wider project will also be drawn out in Delivarable 1.8 Final Business Case Synthesis Report. Detailed reports are included in five annexes as indicated.

Table of contents

Abbreviations and acronyms

Abbreviation / Acronym	Description
TOC	Train Operating Company
MC	Marginal Cost
MCq	Marginal Costs of quality
MC_{jv}^{W}	Marginal costs per ton-km and vehicle type
V_{max}	Maximum velocity
IM	Infrastructure Manager
VAR	Vector autoregressive

1. Introduction

This deliverable draws together a significant body of research undertaken as part of the project.

A quote below is taken from the description of works (DOW):

> *"…However, there is also a need to identify incentive mechanisms that will help deliver these benefits. This can be achieved through ensuring that different players operating in the rail industry have the right incentives to implement the proposed technical solutions. There are a range of incentive mechanisms operating within European rail systems. Of particular importance are the charging mechanisms for use of infrastructure (track access charging regimes), as these affect how much infrastructure managers are paid for usage of the infrastructure. The wider regulatory, organisational, governance and funding arrangements in which infrastructure managers (and other players) are operating is also important.*
>
> *The research frontier is to develop new understanding of the impact of climate and quality (e.g. reliability) on marginal and average costs; develop our understanding of cost variability and relative efficiency by exploiting new datasets that are more closely aligned to business units; and to extend the research on the impact of track access regimes on behaviour to incorporate the wider regulatory frameworks in which railways operate and how these impact on costs and efficiency. In developing all of these research strands best practice approaches from SUSTRAIL will be built into the approach. The proposed incentives research provides the link between technical solutions, implementation and impact."*

The DOW was developed into an ambitious agenda in Deliverable D1.3 Cost Model Development Report, highlighting the following core aspects of research that would be considered:

1. Studying the impact of quality on costs.
2. Studying the impact of climate on costs.
3. Methodological aspects of marginal cost modelling, including Bayesian techniques, dynamic models, and aggregation issues; and
4. Research on incentives for innovation in railways and how different incentive mechanisms operate and interact / contradict.

It was noted in Deliverable D1.3 that in delivering the objectives, prioritisation decisions would be made to place greater emphasis on different aspects depending on data availability and on where most progress could realistically be made (in particular it was considered that new work on 1 and 2 may not be possible simultaneously (with the likely focus being on area 1) and that we would need to focus attention on a sub-set of the methodological aspects in area 3.

As indicated above our work has therefore focussed on the following five areas:

a. Research on incentives for innovation (Deliverable 1.7 Annex 1)

b. The impact of quality on costs (Deliverable 1.7 Annex 2)
c. Methodological aspects of marginal cost modelling: Estimating the marginal cost of different vehicle types on rail infrastructure (Deliverable 1.7 Annex 3)
d. Methodological aspects of marginal cost modelling: Bayesian techniques (Deliverable 1.7 Annex 4)
e. Methodological aspects of marginal cost modelling: Dynamic techniques (Deliverable 1.7 Annex 5)

The policy motivation for developing the research frontier in these areas is clear. The advantages and disadvantages of alternative railway structures have long been the source of considerable debate. EU legislation requires a degree of separation of infrastructure from operations; however, some countries have adopted full, legal separation (e.g. as in Sweden and the UK), whilst others have adopted a holding company structure (most notably in Germany and more recently in France). Different structures may be expected to have differential impacts on rail industry costs and specifically with respect to the research in this project on innovation. Other factors in addition to structure (e.g. economic regulation; the role of government funding and multi-annual agreements; the terms and length of rail franchises) will also have an impact. Better understanding the barriers to and enablers of innovation in complex rail organisational structures is therefore of great policy importance and these issues are explored through several case studies drawn from the countries covered by the NeTIRail-INFRA consortium (Annex 1).

In respect of the cost of quality, railways across Europe face the challenge of how to enhance performance whilst also improving productivity and reducing costs. Therefore, understanding more about the relationship between cost and quality (in our case as measured by delay minutes) is crucial: that is, how much does it cost to reduce delay minutes. This information can be combined with measures of willingness to pay, to determine an optimal level of quality performance which can then be reflected in performance regimes and other regulatory target mechanisms (Annex 2).

Understanding the marginal cost of railway usage and different types of vehicle is also an important area of research from a policy perspective, since this information is needed to set track access charges for rail operators accessing the rail network. Accurate information on marginal cost should incentivise optimal use of the existing rail network – that is, traffic should not run if it cannot cover its marginal wear and tear cost – and differentiation of charges by vehicle type should incentivise the adoption of track friendly vehicles. Infrastructure managers in some countries use econometric methods to set access charges (e.g. in France), whilst others utilise engineering methods. The research reported in this deliverable advances both econometric techniques (annexes 4 and 5) and an approach that seeks to combine econometric and engineering approaches (annex 3).

Below we include an executive summary on the key results from each of the above research areas followed by a short concluding section. Detailed reports are included in five annexes as indicated.

2. Summaries and conclusion of research

2.1 Research on incentives for innovation (Deliverable 1.7 Annex 1)

The NeTIRail-INFRA project is concerned with innovation in the rail sector and in particular in rail infrastructure. This report specifically concerns the incentives for innovation in the European rail industry. European law requires the separation of rail infrastructure from operations, either completely or as separate subsidiaries of the same holding company. It also requires open access for commercial freight and international passenger train operators, and increasingly there is competition within the domestic passenger market, either for public service contracts or purely commercial competition. It also requires a regulator to regulate track access charges and to ensure non-discrimination.

Our literature review found much concern that vertical separation and regulation might discourage investment and innovation, but no agreement on the matter. Within the rail industry there was some evidence of vertical separation increasing costs, through additional transactions costs and non-alignment of incentives.

We used documentary evidence and interviews to review the situation in Britain, Sweden, Germany, France and Slovenia. The most extensive and complicated case study was that for Britain, where a great deal of effort has been put into designing track access charges, performance regimes and other incentives to encourage efficiency and innovation. By contrast in many other countries, including Germany and France, track access charges are basically levied per train km, with inadequate differentiation to give any incentive for the use of track friendly rolling stock, and neither country currently has a performance regime.

It was concluded that there was generally a problem with incentives for efficiency and innovation, arising fundamentally from two sources. Firstly, fragmentation of the structure of the industry meant that innovations paid for by one organisation may produce benefits mainly for others. Secondly, regulatory arrangements, and competitions for franchises where these existed, tended to place emphasis on short run cost savings rather than life cycle costs. On the other hand, where such arrangements did not exist there was a concern that there was inadequate pressure on the infrastructure manager's costs.

Possible solutions were found. For instance, where there remained a dominant operator, the holding company model could provide better aligned incentives, provided that it played an active role in integrating decision taking by its subsidiaries, and might itself play a leading role in encouraging innovation. On the other hand, for the industry to be dominated by an existing incumbent might itself discourage innovation; new entrants might bring new ideas to the industry.

Where franchises cover most of the services run in a particular area, long vertically integrated franchises in which the infrastructure is leased to the train operator might overcome both problems, although there remains an issue about incentives to innovate and invest in the later years of the franchise.

Thus there is no single solution appropriate for all circumstances, but the issue needs careful consideration in each case in the light of the situation it poses.

2.2 The impact of quality on costs (Deliverable 1.7 Annex 2)

Overview

Travel time reliability is a key element of any transport system. An element of quality. In the railway sector, much has been discussed about the costs of delays to passengers and their willingness-to-pay to reduce them, i.e. the demand side of the market. However, delays in the supply side of transport markets have received far less attention (Van Oort, 2016). Similarly, quality in railway cost studies has often been neglected or considered in an ad-hoc basis. This paper fills several gaps in the transport and railway literature by studying the relationship between the costs of railway supply and the degree of travel time reliability. First, we articulate a generic theoretical framework for the relationship between costs and quality. We introduce the notions of marginal proactive cost and marginal reactive cost, leading to a U-shaped relationship between cost and quality. The framework acknowledges that low reliability could be associated with higher (reactive) costs but, similarly, high reliability may need high (proactive) costs too. Secondly, we apply the framework to a dataset of train operating companies (TOCs) in the UK over a period of five years. The estimated cost model allows us to empirically observe the cost-reliability relationship and obtain estimates of the marginal costs of improving reliability. The framework and analysis can be used to aid quality related decisions of TOCs, Infrastructure Managers and regulators in the railway industry. The proposed framework can also be applied to other cost-quality contexts, in and outside transportation.

Key contributions and findings

The paper explores the relationship between travel time reliability and costs of train operating companies. In this process, our work makes a number of additional contributions. First, the paper has brought together several bodies of literature which had remained disconnected from each other. We have uncovered a central theme ‘costs and quality’ that applies across literature 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; management research has widely covered the topic developing Cost of Quality models from 1951, 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 too specific and remained unaware and disconnected from existing Cost of Quality models.

Secondly, we developed a generic 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). In line with the existing literature, the quality-cost relationship is marked by countervailing forces: proactive and reactive costs. Our framework proposed that marginal cost of quality is a combination of marginal proactive costs and marginal reactive costs. The framework was then applied to the context of train operating companies in the UK, but remains general enough to be used in other transport and non-transport contexts.

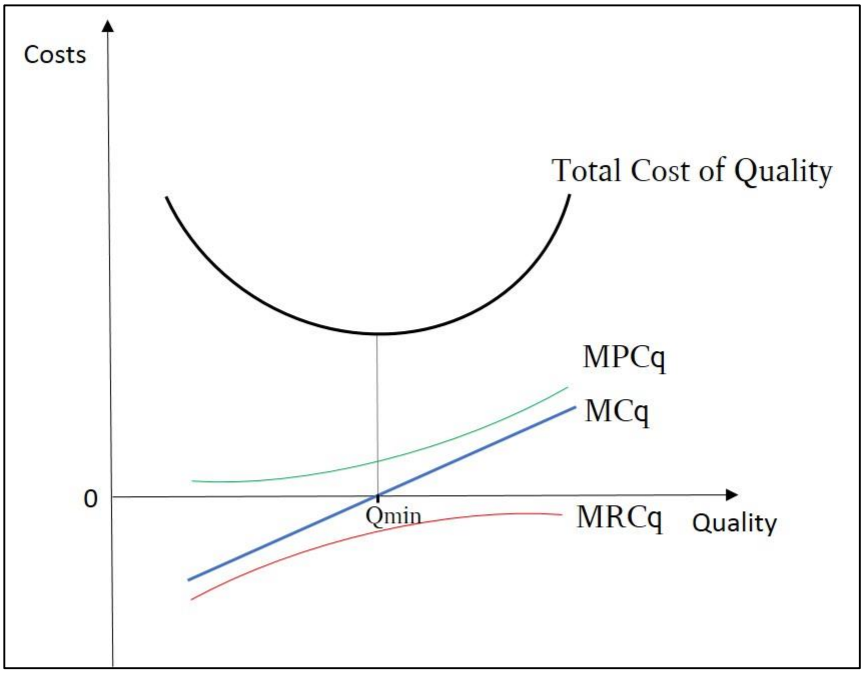


Figure 1 Costs and Quality theoretical framework, from a marginal cost perspective

Third, we provided the first estimates on cost elasticity and marginal costs of reducing delays for train operating companies. The empirical work has shown that in most cases TOCs have been facing a positive marginal cost of quality. At the sample average, we found that it is costly to reduce delays for TOCs. This was expected given that TOCs operate under a performance incentives regime. The cost elasticity of quality was on average around 0.07, but it varied within and across TOCs. TOCs characteristics like average length of trains, vehicle load or salaries were found to influence the estimated elasticities. Also, the marginal cost of quality increased with the level of quality, as theoretically expected. This heterogeneity also revealed that the marginal cost of quality was also not significantly different from zero or negative in approximately 18% of the cases. In those cases, reducing delays would also bring cost savings, and it remains unclear why TOCs would not exploit such opportunity. One possible explanation might be the presence of industry structure constraints that prevent TOCs from doing so. Overall, the results were highly consistent with the theoretical framework.

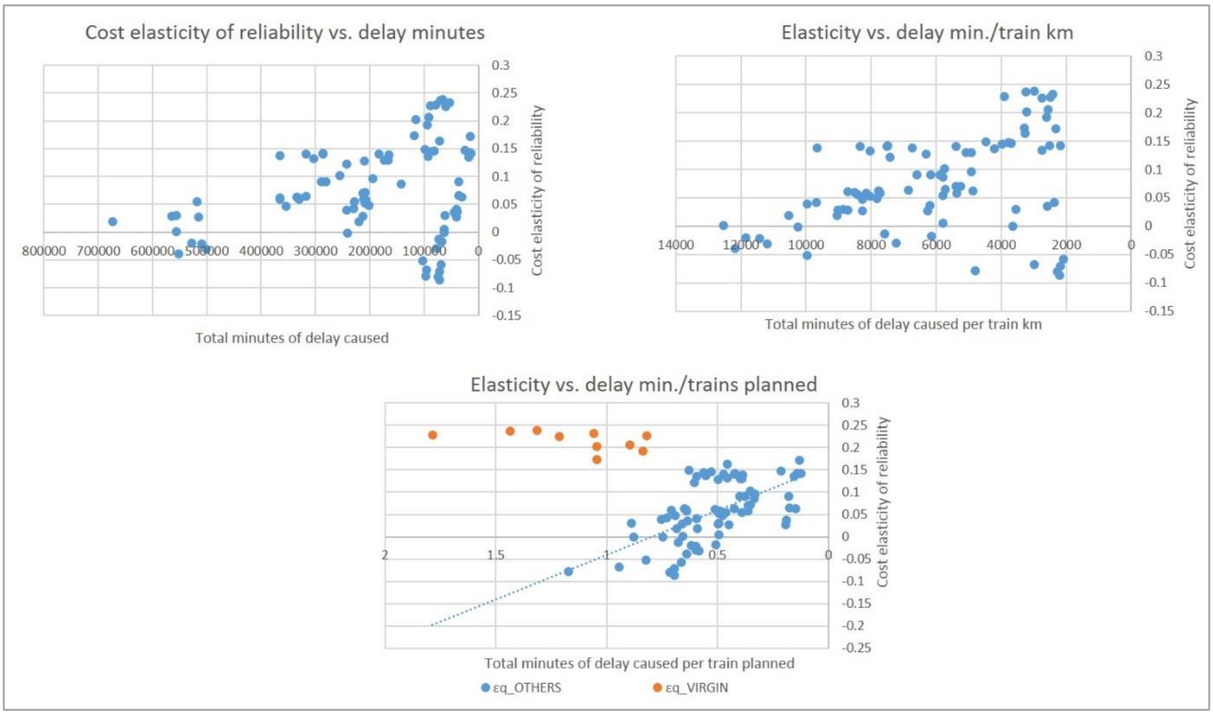


Figure 2 Estimates of cost elasticity of quality

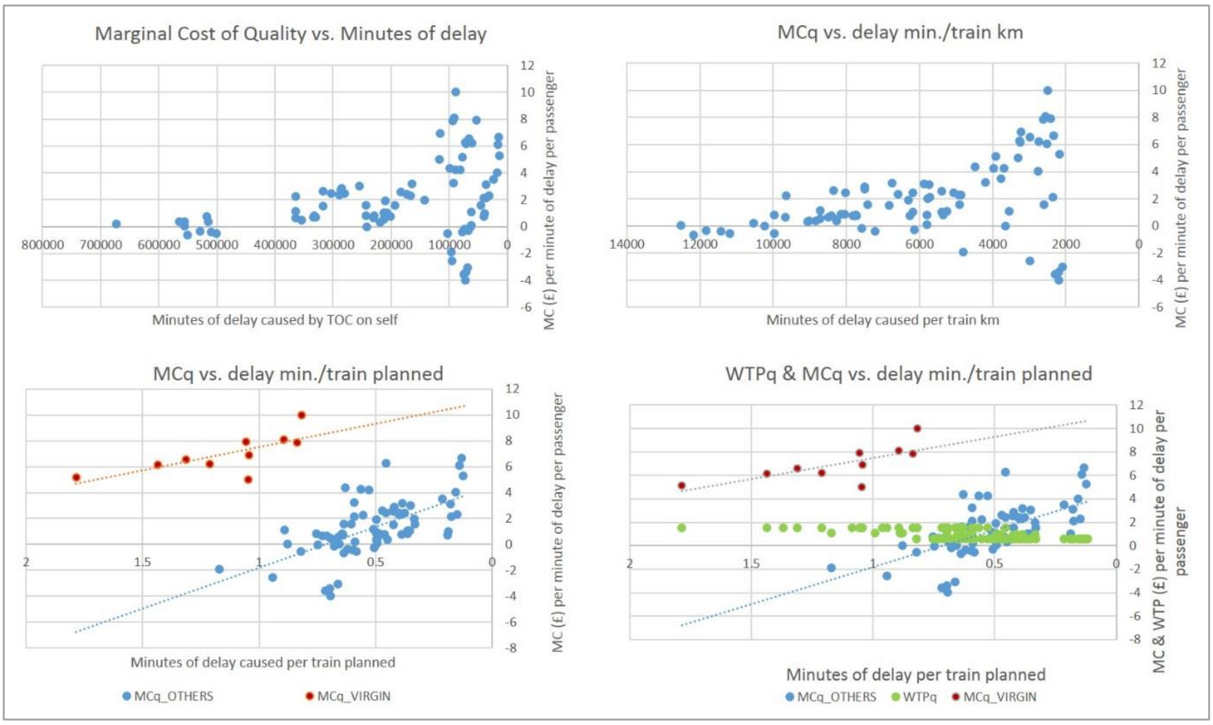


Figure 3 Estimates of Marginal Costs of quality (MCq)

This paper is only a first step to improve our understanding of quality in railway supply. Both the theoretical framework and the empirical application can be used to aid the design of incentives systems and industry structure in relation to quality aspects. Travel time reliability is one aspect of quality, but the framework can easily be translated to other quality contexts. Finally, new empirical evidence would be highly welcome to contrast and complement the first estimates of marginal cost of delay reductions provided in this paper.

2.3 Methodological aspects of marginal cost modelling: Estimating the marginal cost of different vehicle types on rail infrastructure (Deliverable 1.7 Annex 3)

A combination of engineering and economic methods is used to estimate the relative cost of damage mechanisms on the Swedish rail infrastructure and marginal costs of different vehicle types. The former method is good at predicting damage from traffic, while the latter is suitable for establishing a relationship between damage and cost. The best features of both methods are used in a two-stage approach, demonstrating its applicability for rail infrastructure charging.

The estimations are based on 143 track sections comprising about 11 000 km of tracks. In the first stage, simulations based on engineering models are performed to predict the damage caused by different vehicles during 2014. Inputs in the simulations are ideal track geometry and track irregularities, vehicle speeds, wheel and rail profiles, and axle loads. Moreover, vehicle models in the simulations are chosen depending on the traffic that ran on the 143 track sections during 2014. The damage outputs from the simulations are measures on track settlement, wear of rail, rolling contact fatigue (RCF) and track component fatigue. The damage measures are used in the second stage of the approach, in which a statistical model is specified where maintenance cost is a function of the different damage measures as well as other cost drivers. The statistical model is estimated using information on actual costs during 2014, which results in cost elasticities with respect to the damage mechanisms. These elasticities indicate the relative costs of the damages. However, a strong correlation between the damages at the track section level makes it difficult to isolate the cost impact of each damage type. Still, the preferred model provides significant cost elasticities for wear of rails and track settlement. These damage types capture the cost impact from RCF and track component fatigue given the strong correlations between the damages on different track sections.

The cost elasticities for the damage mechanisms are used to derive the marginal cost per damage unit. Together with information on the amount of damage per ton-km each vehicle type has caused, a marginal cost per ton-km and vehicle type is calculated. The results show a substantial variation in the marginal cost per ton-km for different vehicle types running on the Swedish railway. However, this variation is mainly driven by the cost elasticities for wear and track settlement, as well as the differences in wear per ton-km and track settlement per ton-km between the vehicle types. Hence, the marginal costs per vehicle type do not reflect the vehicles’ relative differences in RCF per ton-km and track component fatigue per ton-km.

Table 1 Damages and marginal costs (MC_{jv}^{W}) per ton-km and vehicle type

Vehicle type	Wear per ton-km	Settlement per ton-km	MC wear[a]	MC settlement [a]	Total MC[a]
Motor coach 4x21 t, V_{max} 200 km/h *	209.76	995 468	0.0295	0.0093	0.0389
Three-piece bogie 4x30 t, V_{max} 60 km/h	97.56	867 067	0.0137	0.0081	0.0219
Passenger car 4x14 t, V_{max} 160 km/h	57.34	741 423	0.0081	0.0069	0.0150
Freight loco 6x30 t, V_{max} 70 km/h	36.85	1 001 992	0.0052	0.0094	0.0146
Freight loco 6x20 t, V_{max} 120 km/h	36.90	945 300	0.0052	0.0089	0.0141
Motor coach 4x16 t, V_{max} 200 km/h**	41.46	852 697	0.0058	0.0080	0.0138
Passenger Loco 4x19 t, V_{max} 175 km/h	40.69	740 151	0.0058	0.0070	0.0128
Three-piece bogie 4x6.5 t, V_{max} 60 km/h	50.22	602 992	0.0071	0.0056	0.0127
Passenger Loco 4x19 t, V_{max} 140 km/h	40.85	748 934	0.0057	0.0069	0.0127
Motor coach, Jacob bogie 3x16.5 t, V_{max} 160 km/h**	53.58	476 803	0.0075	0.0045	0.0120
Y25 bogie 4x22 t, V_{max} 100 km/h	30.32	795 901	0.0043	0.0075	0.0117
Freight wagon 2x6.5, V_{max} 100 km/h	49.75	383 151	0.0070	0.0036	0.0106
Motor coach, Jacob bogie 3x12.5 t, V_{max} 200 km/h***	33.73	571 887	0.0048	0.0054	0.0101
Freight loco 4x20 t, V_{max} 120 km/h	21.75	743 656	0.0031	0.0070	0.0100
Motor coach 4x12 t, V_{max} 140 km/h***	21.12	668 032	0.0030	0.0063	0.0092
Freight wagon 2x22 t, V_{max} 100 km/h	26.48	464 017	0.0037	0.0043	0.0081
Motor coach 4x16 t, V_{max} 200 km/h***	12.03	676 894	0.0017	0.0063	0.0080
All vehicles, weighted average (eq. 13)	44.10	751 142	0.0062	0.0700	0.0132

[a] SEK in 2014 prices * High center of gravity and stiff wheelset guidance, ** Stiff wheelset guidance, ***Flexible wheelset guidance

All in all, this study demonstrates how the estimated relative costs of damage mechanisms can be used to calculate the marginal wear and tear cost of different vehicle types. The results are relevant for infrastructure managers in Europe who desire to differentiate their track access charges such that each vehicle pays its short run-marginal wear and tear cost, which can create a more efficient use of the rail infrastructure. More observations over time can be useful in future research in order to provide more reliable and robust estimates, as well as for isolating the cost impact from each damage mechanism.

2.4 Methodological aspects of marginal cost modelling: Bayesian techniques (Deliverable 1.7 Annex 4)

The purpose of this paper is to explore the extent to which a Bayesian approach to rail infrastructure cost modelling can help improve the robustness of such models. The Bayesian approach is expected to be particularly beneficial when relatively small sample sizes can be combined with prior information, obtained from previous cost modelling exercises in comparable contexts, on the parameters of interest. The Bayesian approach is also expected to be able to accommodate the use of flexible econometric specifications whilst maintaining the consistency of the model with economic theory, for instance, by imposing concavity of input prices.

We illustrate our work with an application to SNCF in France.

The overall contribution of this work to the state-of-the-art in empirical analysis of marginal wear and tear costs for rail infrastructure is:

- Demonstration of the approach and feasibility of Bayesian estimation in this context
- Demonstration of the benefit of Bayesian estimation: namely the ability to impose prior information on the implied elasticity relationship to exploit the findings from a large body of empirical studies in this area. This is potentially of great benefit when sample sizes are small
- Consideration of measures to determine whether the data under consideration is compatible with the prior information i.e. are the priors valid?
- Consideration of the benefits of this approach vis-à-vis a classical approach for different sample sizes, showing the importance of prior information for smaller sample sizes.

Approach

Bayesian analysis departs from classical statistical analysis (e.g. Ordinary Least Squares regression), since it assumes that model parameters are essentially random and that their distribution can be learned about through multiple studies which, when combined, produce the most appropriate estimate of the parameter given the sum of the information available at a given point in time.

This is very relevant to the issue of marginal cost for a number of reasons:

- There exists a large number of past studies undertaken across Europe. Further there exists established best practice synthesis such as Wheat et al (2009), which have produced recommendations for the values of the elasticity of cost with respect to traffic.
- Infrastructure managers, overseen by economic regulators, are required by EU legislation to set their track access charges based on an estimate of the marginal cost of running extra vehicles on the network. The legislation provides for different methods to be used, including econometric methods. A key question for infrastructure managers and regulators alike is first of all whether to undertake a new econometric study using their own country data or to rely entirely on past estimates. An econometric exercise is non-trivial in terms of data collection (particularly if the desire is to obtain a large enough sample) and also the skills required to undertake the exercise.

- Taking the above two points together, a further question – which is a key focus of this paper – is whether infrastructure managers / regulators could develop their own, bespoke econometric study, whilst utilizing the information contained in previous work in the form of informed priors within a Bayesian framework. Such an approach could be particularly valuable in a situation where the bespoke dataset is small. Of course the necessary econometric skills would need to be available or bought in to implement such an approach.

In Bayesian analysis a parameter has a 'prior' distribution, which embodies what is known about the parameter before undertaking the study in question. In this paper, we use the Wheat et al (2009) recommendations for the range of appropriate values for the cost elasticity with respect to traffic as the prior information and then derive, through Bayesian estimation, the posterior distribution and thus an estimate using various splicing of a dataset from France. We consider three prior scenarios as shown in Table 2.

Table 2 Relationship between prior distribution and estimates from the posterior distribution

Prior Distribution Class	Implementation in Section 5	Properties of estimates from Posterior Distribution
Non-informative prior	Normal distribution with very large variance	Parameter estimates will be very similar to those from classical statistics e.g. OLS
Informative prior	Normal distribution N(0.2,0.01)	Parameter estimates are not forced to be within any bound however the prior will influence the estimates to the extent that the estimates will likely be closer to the prior mean relative to the OLS estimates. As the sample size increases, the influence of the prior diminishes, such that the estimates will approach the results from classical statistics
Informative prior – bounded parameter space	Uniform distribution. In section 5, we assume the traffic elasticity is Uniform[0.2,0.35]	Parameter estimates are forced within the bounds of the prior density ([0.2, 0.35] in section 5).

Findings

Our key conclusions are:

- We have been able to estimate Bayesian formulations of infrastructure cost functions for a dataset for France. Estimation results are comparable to those from classical approaches, however there is a clear influence of the prior information as intended.

- We have demonstrated that the influence of prior information for the posterior estimates is most influential when there is a limited sample size. This indicates that Bayesian analysis might be very beneficial when sample sizes are limited. Indeed, these techniques could be of great benefit where a country is considering developing an econometric study of marginal wear and tear costs for the first time (as the data requirements are less than trivial for a full classical study). Of course this assumes that the analyst and policy maker has confidence in the appropriateness of the prior information.

- We have outlined a measure of prior data conflict which highlights when the prior information is incompatible with the data in the sample. This is important for determining the appropriateness of the prior. Applying this criterion to our dataset reveals that bounded priors (fixed ranges of permissible values for an elasticity in our application) do lead to instances of prior data conflict. However given the nature of the generalisation framework in Wheat et al (2009) which involved judgement over a wide range of studies, unbounded priors are most appropriate.

This work represents a first exploration of the value of applying Bayesian techniques to the problem of marginal wear and tear cost estimation. Recommendations from this and subsequent work would hope to inform whether a infrastructure manager or economic regulator should rely only on their own available dataset to inform charges, or whether they should be explicitly drawing on past information, such as from the FP7 CATRIN project (Wheat et al, 2009), either as prior information in Bayesian econometric work or relying on past evidence only.

We consider that a natural extension of this work is to generalise the techniques to impose prior information on elasticities in second order cost functions e.g. Translog as these provide a more flexible descriptions of costs and represent the state-of-the-art. This involves assigning priors to both functions of model parameters and data, which is a subject for further research.

2.5 Methodological aspects of marginal cost modelling: Dynamic techniques (Deliverable 1.7 Annex 5)

The purpose of this paper is to provide empirical evidence on the wear and tear costs of rail infrastructure in Sweden, using a 16 year panel dataset. To do this we consider a dynamic panel data specification which allows for interdependence between maintenance and renewals, as well as their intertemporal effects. The estimates can be used to calculate the marginal cost for traffic, which has become an important part of the track access charges that were introduced after the vertical separation between train operations and infrastructure management in Europe as of the 1990s. Given that there are dynamic effects between different activities in infrastructure provision, the marginal cost estimates that takes these effects into account will be closer to the actual cost of running one extra unit of traffic on the railway, compared to the cost estimates based on static models for maintenance (see for example Wheat et al. 2009) and renewals (see for example Andersson et al. 2012 and Andersson and Björklund 2012).

Methodology

In this paper we analyze the dynamics between rail infrastructure renewals and maintenance in Sweden, using a panel vector autoregressive model. The dynamics in maintenance and renewals implies that an infrastructure manager (IM) needs to strike a balance within and between these activities for a certain traffic level. A sudden increase in traffic may thus require an adjustment of these costs. This implies that the cost impact from traffic needs to be studied in a dynamic context. The model estimation also comprises intertemporal effects for each of these activities.

We consider a panel VAR(p) model, where p denotes the lag length used in the model.[1] We have two endogenous variables: renewal costs (R_{it}) and maintenance costs (M_{it}), where $i = 1,2 \dots, N$ contract areas and $t = 1,2, \dots, T$ years. $\alpha_{1,i}$ and $\alpha_{2,i}$ are the unobserved individual-specific effects for the

[1] Here we present the VAR(1) model for expositional simplicity. We consider further lags in the model estimation.

renewal and maintenance equations respectively, while $u_{1,it}$ and $u_{2,it}$ are their respective residuals, where $\left(u'_{1,it}, u_{2,it}\right) = \boldsymbol{u}_{it} \sim iid(0, \sum)$. $\sum$ is the covariance matrix of the errors. We also include a vector of exogenous variables $\boldsymbol{X}_{it}$ with parameters $\boldsymbol{\beta}_{11}$ and $\boldsymbol{\beta}_{21}$ for the maintenance and renewal equations respectively.

$$R_{it} = \alpha_{1,i} + \delta_{11} R_{it-1} + \theta_{11} M_{it-1} + \boldsymbol{\beta}_{11} \boldsymbol{X}_{it} + u_{1,it}$$

$$M_{it} = \alpha_{2,i} + \delta_{21} R_{it-1} + \theta_{21} M_{it-1} + \boldsymbol{\beta}_{21} \boldsymbol{X}_{it} + u_{2,it} \qquad (1)$$

Lagged renewal and maintenance costs are included in both equations to capture the dynamics in maintenance and renewals, as well as the interdependence between these activities.

Data

We have cost data for renewals and maintenance cost separately. That helps us build our two equation model in (1).

A full set of variables available for the analysis is given in Table 3.

Table 3 – Descriptive statistics, 1999-2014 (480 obs.)

	Mean	St.dev.	Min	Max
Hourly wage, SEK*	156.7	11.7	128.9	187.4
MaintC (Maintenance costs), million SEK*	56.78	44.37	8.03	334.41
RenwC (Renewal costs), million SEK*	40.74	63.95	0.00	452.13
Route length, km	280	174	13	989
Track length, km	358	229	39	1203
Length of switches, km	8.68	6.62	0.58	37.67
Length of structures (tunnels and bridges), km	5.72	7.22	0.55	40.43
Average age of rails	18.83	5.83	3.76	38.98
Ton-density (ton-km/route-km), million	7.9	7.2	0.2	33.2
Mixtend	0.06	0.24	0	1
Ctend	0.47	0.50	0	1
Trend	8.45	4.50	1	16

* 2014 prices.

Results

We find that past values of maintenance gives a better prediction of current renewal costs compared to only using past values of renewals as a predictor. That is, we find evidence for dynamic effects which is in turn an endorsement for our approach. Moreover, the results indicate intertemporal effects for both renewals and maintenance, where an increase in costs during a year predicts an increase in costs in the following year.

A particular purpose of estimating the dynamics in infrastructure costs is that it allows us to take these effects into account when assessing the cost impact of traffic. The estimated cost elasticity with respect to traffic is different in our dynamic model compared to static models that are frequently used in the literature on rail infrastructure costs. In particular, we used information on the dynamics between renewals and maintenance, in order to estimate equilibrium cost elasticities. These elasticities can be used in the calculation of marginal cost (which is the product of average costs and the cost elasticity), giving a better representation of the cost impact of an additional ton-km, considering that a traffic increase gives rise to costs in both the current year and subsequent years. Hence, the results in this paper are informative for infrastructure managers in Europe who need to set track access charges for the wear and tear caused by traffic.

Key elasticity results are shown inTable 4 . The parameter estimate for ton density in the maintenance equation is 0.2330 (p-value=0.010), which is in line with previous results on Swedish data (see for example Odolinski and Nilsson 2017 or Andersson 2008). In the renewal equation, the coefficient for ton density is 0.2633, which is lower than previous estimates on Swedish data (however, our estimate is not significantly different from zero, p-value = 0.498); Andersson et al. (2012) find a cost elasticity with respect to ton density at 0.547, and Yarmukhamedov et al. (2016) find elasticities between 0.5258 and 0.5646.

We calculate the equilibrium cost elasticities with respect to ton-density for both renewals and maintenance, using the results from model A2. These are presented in Table 4, where γ^e denotes equilibrium cost elasticity. The elasticity for renewals is not significant at the 10 per cent level, while the estimates for maintenance are significant at the 1 per cent and 5 per cent level. All in all, the elasticities are larger than their static counterparts.

Table 4 Key elasticity results from this study

	Cost elasticity	Coef.	Std. Err.
Dynamic Short run elasticity	$\gamma_{Maintenance}$	0.2330***	0.0901
	$\gamma_{Renewals}$	0.2633	0.3885
Dynamic Equilibrium elasticity	$\gamma^e_{Maintenance}$	0.3376**	0.1352
	$\gamma^e_{Renewals}$	0.3439	0.5173
Static comparator model	$\gamma_{Maintenance}$	0.2431***	0.0769
	$\gamma_{Renewals}$	-0.1189	0.4612

	$\gamma_{Main+Ren}$	0.2506**	0.1197

Notes: Short run elasticity is the change in cost resulting from a change in traffic in the same period; Equilibrium elasticity is the change in cost resulting from a change in traffic once all of the impact has traced through over time. Static comparator models represent results from more conventional contemporaneous models (they do not have dynamic terms included).

Overall, this work highlights that the dynamics in rail infrastructure costs are important to consider when setting track access charges with respect to the wear and tear caused by traffic.

The results can also be a useful demonstration of the maintenance and renewal strategy currently used. For example, the estimate for the second order lag of maintenance cost in the renewal equation gives us a hint on how sensitive renewal costs are to prior increases in maintenance. Moreover, the intertemporal effect for maintenance reveals how quickly this cost adjusts to equilibrium. Still, there is more to be done in this research area. For example, the analysis in this paper is not able to answer whether the quick adjustment in maintenance costs is avoiding an over-investment - that is, doing more than is necessary to uphold the performance of the infrastructure. In fact, the IM may well be over- or under-investing in maintenance after a sudden increase in traffic. User costs (values of train delays for passengers and freight companies) must be considered in this type of analysis. That is, with access to data on train delaying failures and delay costs for passengers and freight companies, it could be a step towards a cost-benefit analysis of maintenance and renewals which in turn can generate economically efficient levels of these activities. This is an area for future research.

3. Concluding remarks

The research outlined in this deliverable has developed the research frontier in several areas with results that are both generalizable and have specific relevance to the case study countries and to the innovations being developed in the other work packages. In Deliverable 1.4 Cost and User Benefit Report the costs and the benefits of the NeTIRail-INFRA project will be established. Deliverable 1.8 Final Business Case Synthesis Report will pull together the findings of the present deliverable (Deliverable 1.7 Incentives Final Report) and Deliverable 1.4, setting out the overall business case for the innovations, considered alongside the incentives for their implementation.

Pending the integration of Deliverable 1.7 with Deliverable 1.4, some brief comments are made here concerning the specific implications of the research in Deliverable 1.7 for the case studies considered in this project.

In respect of Annex 1, two primary barriers to innovation and efficiency were found to be: (1) fragmentation, leading to a misalignment of incentives between different parts of the industry where the costs may be carried by one party and the benefits felt by another; and (2) regulatory and franchising arrangements leading to a short-term focus in place of a focus on life cycle costs. Given that EU countries are required to introduce competition for passenger services, which for public service contracts will mean increased adoption of franchising, there are important lessons here for the case study countries.

Whilst the introduction of competition can lead to cost reduction and potentially innovation, short-franchises, combined with a vertically-separated rail infrastructure manager facing regulatory targets, can cause significant challenges for co-ordination of innovation and investment and a focus on short-term cost reductions at the expense of sensible long-term planning. Funding constraints imposed by government can also have this effect, making it more cost effective in the short-term to favour maintenance over renewals. In principle, an independent economic regulator could play an important role here in terms of ensuring a longer-term focus and potentially encouraging co-ordination, but there may be limits to what can be achieved in practice. Of the three case study countries in this project, only Romania has a vertically separated structure.

Importantly, contractual mechanisms (track access charges and performance regimes, if calibrated correctly, can play an important role in aligning incentives within a separated structure. However, in general across Europe there is a lack of differentiation of track access charges for different vehicle types (Britain being an exception). Greater differentiation of track access charges should encourage the development of vehicle designs that do less track damage as the incentives of operators would be more closely aligned with the infrastructure manager.

As an alternative to vertical separation, many countries across Europe have adopted a holding company model, including Germany, France and Slovenia (Turkey also has a structure similar to the holding model). Through the interviews conducted as part of the research reported in Annex 1, it is clear that the holding model can offer solutions to the co-ordination in principle and in practice (if the holding plays a significant co-ordination role); however this model could also reduce the extent of new entry, which in turn could hamper innovation.

Turning to the research contained in the other annexes of this deliverable, Annex 3 is specifically concerned with developing a new method for estimating the relative cost of damage imposed by different rail vehicles. Such research is important because it is not sufficient to recognise that track access charges should be differentiated by vehicle type it is also necessary to calculate those relative costs. Annex 4 is concerned more with the general variability of rail infrastructure costs with respect to traffic (needed to calibrate track access charges) and offers a means of obtaining better estimates when an individual country has limited data available for estimation. Such an approach could be useful for the case study countries where obtaining disaggregate data has proved to be a challenge. Annex 5 likewise proposes methodological enhancements to better estimate marginal costs of rail infrastructure usage, with a view to developing more cost reflective access charges. Strengthening the evidence base in this area is important for EU countries (given the legislation), but also to any railway seeking to set access charges based upon marginal wear and tear costs (and indeed to Turkey which seeks to align its policy with EU legislation).

Finally, Annex 2 focuses on a particular aspect of cost modelling that is relevant to many of the innovations – namely the relationship between cost and quality. It may not be possible to directly apply this analysis to the case study countries because of a lack of data available on the relevant quality metrics (e.g. delay minutes for the case study lines). However, this work points not only to the idea that improving quality is likely to require increased preventative costs, but also to the possibility of a lose-lose scenario, where quality is poor and reactive costs are so high as to lead to a situation where quality is low and overall costs are higher than they need to be. Similar relationships have been observed in the health sector.

The relationship between the work in this deliverable and the final business case will be considered further in Deliverable 1.8 as noted above.

4. References

See individual annexes for references.

Funded under the H2020 programme

Collaborative project H2020-MG-2015-2015 GA-636237

Needs Tailored Interoperable Railway – NeTIRail-INFRA

Deliverable D1.7
Incentives Final Report Annex 1
Research on Incentives for Innovation

Document ID: NeTIRail-WP1-D1.7v1.0-FINAL – ANNEX1

Due date of Deliverable: 30/09/2017

Actual submission date: 21/12/2017

Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Task leader for this deliverable: Professor Andrew Smith, Institute for Transport Studies, University of Leeds

Document status		
Revision	Date	Description
V0.1		First draft (authors Chris Nash, Bryan Matthews and Andrew Smith)
V0.2	14/12/2017	Review by USFD and ALU-FR
V1.0	21/12/2017	Final version
Reviewed	YES	

Executive Summary

The NeTIRail-INFRA project is concerned with innovation in the rail sector and in particular in rail infrastructure. This report specifically concerns the incentives for innovation in the European rail industry. European law requires the separation of rail infrastructure from operations, either completely or as separate subsidiaries of the same holding company. It also requires open access for commercial freight and international passenger train operators, and increasingly there is competition within the domestic passenger market, either for public service contracts or purely commercial competition. It also requires a regulator to regulate track access charges and to ensure non-discrimination.

Our literature review found much concern that vertical separation and regulation might discourage investment and innovation, but no agreement on the matter. Within the rail industry there was some evidence of vertical separation increasing costs, through additional transactions costs and non-alignment of incentives.

We used documentary evidence and interviews to review the situation in Britain, Sweden, Germany, France and Slovenia. The most extensive and complicated case study was that for Britain, where a great deal of effort has been put into designing track access charges, performance regimes and other incentives to encourage efficiency and innovation. By contrast in many other countries, including Germany and France, track access charges are basically levied per train km, with inadequate differentiation to give any incentive for the use of track friendly rolling stock, and neither country currently has a performance regime.

It was concluded that there was generally a problem with incentives for efficiency and innovation, arising fundamentally from two sources. Firstly, fragmentation of the structure of the industry meant that innovations paid for by one organisation may produce benefits mainly for others. Secondly, regulatory arrangements, and competitions for franchises where these existed, tended to place emphasis on short run cost savings rather than life cycle costs. On the other hand, where such arrangements did not exist there was a concern that there was inadequate pressure on the infrastructure manager’s costs.

Possible solutions were found. For instance, where there remained a dominant operator, the holding company model could provide better aligned incentives, provided that it played an active role in integrating decision taking by its subsidiaries, and might itself play a leading role in encouraging innovation. On the other hand, for the industry to be dominated by an existing incumbent might itself discourage innovation; new entrants might bring new ideas to the industry.

Where franchises cover most of the services run in a particular area, long vertically integrated franchises in which the infrastructure is leased to the train operator might overcome both problems, although there remains an issue about incentives to innovate and invest in the later years of the franchise.

Thus there is no single solution appropriate for all circumstances, but the issue needs careful consideration in each case in the light of the situation it poses.

Table of contents

Abbreviations and acronyms

Abbreviation / Acronym	Description
SERA directive	Single European Railway Area; Directive 2012/34/EU
IM	Infrastructure Manager
BC	Benefit Cost
DBB	Design-Bid-Build
BoQ	Bill of Quantities
DB	Design-Build
PPP	Public Private Partnerships
R&I	Research and Innovation
EU	European Union
R&D	Research and Development
EC	European Commission
OECD	Organisation for Economic Co-operation and Development
PR	[Network Rail 5 yearly] Periodic Review
CP	[Network Rail] Control Period
HLOS	[UK government] High Level Output Specification
SOFA	[UK government] Statement of Funds Available
ORR	[UK] Office for Rail and Road
ETCS	European Train Control System
RAB	Regulatory Asset Base
DfT	[UK] Department for Transport
TOC	Train Operating Company
RDG	[UK] Rail Delivery Group
REBS	Route-Level Efficiency Benefit Sharing Mechanism
IEP	[UK] Intercity Express Programme
RSSB	[UK] Rail Safety and Standards Board
ERTMS	European Railway Traffic Management System
PTE	Passenger Transport Executive
NR	Network Rail
CBA	Cost Benefit Analysis
BC	Benefit Cost test
B2B	Business to Business
B2C	Business to Consumer
ABA	Axle Box Acceleration monitoring
GPS	Global Positioning System
DB	Design Build
DBB	Design Bid Build
DB Netze	Deutsche Bahn Netz (IM)
ICE	[Deutsche Bahn] Intercity Express
PPP	Public Private Partnership

1. Introduction

1.1 Introduction

The SERA Directive (a Single European Railway Area; Directive 2012/34/EU) establishes a common European approach to organising its railway industry. The separation of railway traffic operations from the provision of infrastructure is at the core of this approach, but within the overall setting of the Directive, there is a substantial leeway for member countries to organise its industry. In practice, the main alternatives are complete vertical separation or retaining infrastructure and the main operator as separate subsidiaries of a single holding company.

One part of the Union's objectives when drafting the SERA Directive is to contribute to the efficient use of resources in the railway industry. The present inquiry addresses one dimension of this objective, commonly referred to as dynamic efficiency which inter alia is concerned with the industry's innovativeness. An overriding question is therefore whether the way in which the EU has organised its railway services makes the market innovative.

NeTIRail-INFRA, a project funded by the European Commission, provides the trigger for the discussion. NeTIRail-INFRA is developing proposals for 10 innovations in the way that railway infrastructure is monitored and maintained. One part of the project is to make an economic assessment of the benefits and costs of each of these innovations. The question is if one, two or indeed all NeTIRail-INFRA's proposals generate more benefits (lower life-cycle costs) than the upfront cost for implementing the respective proposals. If the answer is affirmative, this will be referred to as they pass the BC (Benefit-Cost) test.[1]

The focus of the paper is on innovations with focus on railway track investment and maintenance. In the same way as for any type of investment, an innovation – vaguely defined as a new way or design of doing something – should be implemented if the higher cost 'today' is balanced with lower costs and/or higher benefits 'tomorrow'. While there is a range of explanations of why apparently beneficial changes are not realized, the focus here is on the significance of the institutional setting of the railway industry for undertaking (or not) beneficial projects: Does the institutional framework for organising and regulating the European railway industry provide incentives for dynamic cost efficiency?

Three examples of innovations can provide a background for the discussion.

- An Infrastructure Manager (IM) has come to realise that the present value of the benefits (lower maintenance costs of tracks and rolling stock) of installing heavier rails exceed the extra cost for buying the more expensive rails. Will the way in which the industry is organised result in a decision to upgrade track weight?
- Assume that frequently updated information about track quality can be made available at low cost. This information will increase the chance of early detection of track irregularities triggering fast remedial action at lower cost than if the irregularity had developed into a

[1] Since NeTIRail addresses track maintenance and investment issues, this means that the pros and cons for train operators of vertical separation and its impact on dynamic efficiency is not covered in this mimeo.

problem affecting safety. Spending on information acquisition meets the BC criterion. Will the way in which the industry is organised trigger this change of policy?

- A new type of switch that will reduce the risk of failures during bad winter conditions is on the drawing board but is yet not available in a prototype version. The new (costlier) switch may reduce the IM's costs for future maintenance and operations and the risk for trains to get stuck because of malfunctioning switches. It is not yet clear if the switch meets the BC criterion. Will the potential innovation be taken further from the drawing board to building a prototype?

1.2 The process of infrastructure investment and maintenance in a vertically separated railway industry

Commercial organisations should invest whenever the expected benefits in the form of (the present value of) future increases in (net) income exceed the costs for a project, i.e. if the project meet the BC-test or rule. The profit maximisation hypothesis provides the motive for commercially viable investments – i.e. projects that pass the test – to be implemented. Not only owners but also the firms' managers benefit from the fact that these projects generate higher profits, the latter by way of wage increases and other more indirect benefits.

Most if not all commercial activities have systems for cost monitoring. One motive is that real-time monitoring of the performance of projects that are being implemented increases the chance for that those projects – that presumably have passed the BC test – deliver the (net) benefits originally perceived. Another motive for monitoring is that information about historical project costs and revenue can be used as an input for the assessment of future investment decisions of similar nature.

Applying the BC test in the public sector presupposes a definition of costs and benefits that include all consequences of a measure, irrespective of if they occur within or outside the organisation. Another difference is that the public sector is not supposed to maximise profits but rather the welfare of those affected by a project. This means that the incentives for private owners and their managers to apply the BC test is not automatically transferrable to their public-sector peers.

The purpose here is not to consider the precise motives that may drive the way in which public sector agencies operate to maximize social welfare. Rather, and as indicated in section 1.1, the aim is to consider the implications of rules that provides the governance framework for the railway sector, i.e. the SERA Directive, for dynamic efficiency. For this purpose, section 1.3 reflects over the issue of innovation incentives when comparing a vertically integrated railway sector (the "old approach") to an organisation with separation of infrastructure and train operations. Section 1.4 addresses the Swedish quirk to the vertical separation, i.e. the transfer from in-house to tendered construction and maintenance, and its consequences for dynamic efficiency.

1.3 Innovation in a vertically separated industry

Railways have for years provided a textbook example of an industry with scale economies. It is therefore important to consider whether any concerns with respect to dynamic efficiency is related to the idiosyncrasies of the sector or to the way in which is organised within the European Union context.

Daily production activities are in most industries concerned with delivering products that (historical) investments have made it possible to produce. Train services is the output from the railway industry. To be able to deliver these services, it is inter alia necessary to ascertain that the existing infrastructure is functional. In contrast to daily service delivery, railway investments are project-based and each project is implemented separately from day-to-day activities. It is obvious that the more investment projects under way or on the drawing board, the heavier investment activity becomes relative to everyday service delivery for the industry.

It has already been established that investment decisions are taken with a life cycle perspective; extra spending today will reduce costs for maintenance and/or will provide capacity for more production – more traffic – and extra revenue tomorrow.[2] Maintenance has a shorter perspective in that it is supposed to facilitate ongoing traffic. For both track investment and maintenance, dynamic efficiency is concerned with developing ways and means for reducing the costs for investment and maintenance and/or with developing products that are beneficial for users.

In many member countries, track user charges paid by train service operators do not pay the full cost of their use of the infrastructure. In combination with vertical separation, this severs the link between those that pay the bill for investment and maintenance – ultimately the tax payers – and the train operators. The IM's challenge in a vertically separate railway industry is therefore to account for user benefits and costs which in an integrated railway are represented by in-house costs and benefits. The IM must also consider consequences for the users when choosing between alternative investment proposals as well as when establishing the projects' design. In neither estimate, user benefits and costs appear in publicly available accounting reports. This generates a challenge to acquire relevant proxies for the effects. Even if this is in the instruction to the IM, this does not guarantee that these proxies rather than other concerns are the most important drivers of the trade-offs made by the IM.

This provides a broad characterisation of the challenges that the IM must handle in a vertically separated industry. The aspect that is in focus here, the dynamic efficiency in the way in which the infrastructure is built and maintained, provides an extra challenge in that it is not even in principle possible to recognise all potential improvements that are available.

1.4 Innovation under competitive procurement

As well as implementing the SERA legislation, many European railways have also transferred service delivery from its own staff to commercial firms after competitive tendering of investment and maintenance contracts. The IM is no longer responsible for any delivery of investment or maintenance activities; no agency employees "use shovels for digging" but buy all work is done by entrepreneurs. It is therefore reason to consider what this means for dynamic efficiency in the industry.

[2] Even if welfare concerns should also include positive or negative effects of investment and maintenance interventions for externalities both within the industry and for other modes of transport, this is in practice of second order relevance, at least in a country like Sweden with little road congestion and small cross price elasticity.

The trigger of an investment project is typically a capacity shortage which is addressed by way of building more infrastructure. After a lengthy process, a reasonably precise description of the project this provides a platform for preparing detailed drawings and planning of the production process. This is the blue-print for the work with excavation and refill and for installing the railway specific installation on an embankment.

Even if entrepreneurs and consultants may be involved in the initial part of the planning process, this is still first and foremost a responsibility for the IM also under a policy of competitive procurement. There are then several alternative ways to tender production.

- Design-Bid-Build (DBB): This approach means that the IM first tenders a projector and then – based on the drawings and the Bill of Quantities (BoQ) prepared by the projector consultant – tenders the construction. A bid comprises a unit price for each quantity specified in the BoQ, and the total value of the bid is the product of prices and quantities; in economic jargon, this is referred to as a Unit Price Contract.
- Design-Build (DB): Based on a description of what result that the IM wants, several entrepreneurs prepare a preliminary BoQ of their own and submits an aggregate bid for the project; the subsequent contract with the lowest bidder is typically remunerated by way of a fixed price equal to the winning bid.

The tendering of track maintenance services is also preceded by a Quote for Bids which typically includes a BoQ as a core component. The low-cost bidder is then made responsible for a section of the network under a five- to seven-year period. More information about the design of these contracts is provided in section 4.

The focus here is on the consequences of competitive tendering of railway investment and maintenance for dynamic efficiency. One way to approach this issue is to consider what can be the most extreme version of competitive tendering referred to as Public Private Partnerships (PPP). This is a long-term contract where a commercial firm is made responsible for construction and maintenance of railway infrastructure as well as for operating train services over a longer period.[3] It is still not a shift to full privatisation since the asset will be transferred to the IM after the conclusion of the contract.

In the same way as under the traditional vertically integrated industry structure, a PPP gives most control over, as well as incentives for innovativeness to the contactor. This prevails with respect to both how the new line is designed as well as to the strategy for providing the train service, of its marketing etc. The concessionaire however has no interest in general development of the industry, which includes basic research into equipment and methods that is not of immediate relevance for the project. This obligation remains with the IM.

Contracting based on a DBB contract increases the scope of the IM's responsibility relative to a PPP. The BoQ prepared by the IM or by a consultant on behalf of the IM, describes at great length what the winning entrepreneur is supposed to do and quantifies the extent of each activity. This is one means for ascertaining that a project is implemented in precisely the way in which the IM believes to be relevant. This includes concerns over the life length of the new line. By defining precisely how a

[3] The single Swedish example is the Arlanda railway line, linking the airport to downtown Stockholm. The concessionaire has this contract for a period of 45 years of service; cf. further Nilsson et al (200x).

railway embankment is to be built and transferring this design to the BoQ, and if the entrepreneur implements all activities specified in the BoQ, the risk that the new line will collapse or deteriorate in quality too fast is reduced.

The entrepreneurs that submit bids for the project benefit from being cheaper – having a lower unit price – than their competitors. The competitive pressure suggests that smarter production methods, more productive plant etc. is being developed, providing the outlet for the ingenuity of the entrepreneurs and delivering the prime benefit for society at large from using DBB in competitive procurement. Except for this dimension, all pressure on being innovative related to project design is with the IM. In particular, the designated entrepreneur is not allowed to build in any other way than defined in the BoQ even if this could be demonstrated to be more efficient. This is one basic idea of the regulatory framework for competitive procurement, i.e. that all bidders shall have equal opportunity to develop their own approach for implementing the project.

The use of DB eliminates this brake on innovativeness and provides a middle road for increasing dynamic efficiency. Under this type of tender, bidders are incentivised to look for the smartest – the cheapest – means for reacting to the Quote for Bids sent out by the IM. This could, for instance, mean that the precise alignment of the new line could be adjusted, at least at the margin. Another example that has recently been implemented concerns the way in which an outdated bridge was upgraded. Rather than closing two out of four lanes during the construction period, a new bridge was built next to the existing and then gradually pushed in place over a period of a few days. This was both cheaper and resulted in less disturbance for traffic than traditional methods for bridge renewal.

One downside of the DB contract from the perspective of bidders is that it transfers risk from the IM. This will, all other things equal, induce the bidders to include a risk premium in order to protect from negative risk realisations. While DB increases the scope for being innovative, it also introduces the possibility that the builder does not account for quality in the way that the IM would prefer. To handle this challenge, there are examples of that the entrepreneur is made responsible for the maintenance of the new asset for a sequence of years. Since the entrepreneur is made responsible for the consequences of poor quality, this is a means for increasing the possibility that the DB contract delivers a product which is robust over time.

From the perspective of entrepreneurs and consultants, there are additional features of the railway sector that makes innovativeness less attractive than in many other sectors of the economy. One reason is that innovations often are more related to methods than to new equipment. This affects incentives since it is far more difficult to patent a method than new kit. The example given above of the new way to build the bridge can obviously be used by any entrepreneur that in the future will submit a bid for a similar project.

There may also a built-in brake on developing new equipment in this industry, at least to the extent that this means expensive plant such as purpose-designed trucks, excavators, etc. The reason is that the business cycle in construction may swing more extremely than in many traditional production activities. When demand is low, and in particular under severe depressions, it is feasible to lay off staff to save on costs, while it may be difficult to sell specialised equipment, not least since it is not in demand anywhere. This means that even if resources allocated to developing machinery may per se pass the BC test, entrepreneurs may hold back on these expenses knowing that it would represent sunk costs during downward swings of the business cycle. This is one possible reason for that the construction industry is less productive than the economy at large; cf. further The Economist (2017).

There are more aspects of relevance to incentives for dynamic efficiency by entrepreneurs. An important conclusion is, however, that using the market forces for reducing the costs for delivering investment and maintenance activities over time is no panacea for enhancing the dynamic efficiency of the industry. While a transfer to competitive tendering opens the possibility for using the market also for handling dynamic efficiency, the challenge provided by split incentives in a vertically separated railway industry remains in many while not all efficiency dimensions.

1.5 Overview of the report

The reforms to Europe's railways since the mid-1990s have been aimed at revitalising the performance of railways across Europe through enhanced within-mode competition. To facilitate this, some form of vertical separation is required by law, either full, legal vertical integration (for example, as in the UK or Sweden), or a holding company model (for example, as in Germany and France). As competition has increased the number of companies operating across Europe's railway systems has also increased both in terms of open access train operators, and also train operators providing services through franchise contracts for a fixed number of years. With this, concerns have emerged regarding institutional arrangements and coordination within the industry. Performance regimes and other contractual mechanisms such as track access charges play an important role in facilitating coordination, and more recently intermediate organisational structures, such as alliances have also emerged to encourage co-ordination between companies in separated environments

At the same time, European and country-specific rail research is targeted at bringing about a step change reduction in overall rail system costs. The Shift2Rail Joint Undertaking, set up to fund and develop European rail research over the ten-year period from 2014-2024, therefore faces the challenge of developing new innovations that will meet the cost and other challenges (e.g. on capacity, carbon and customer service).

When it comes to promoting and implementing innovation, there is a concern that fragmentation in rail presents problems, and a priori it seems unclear what structure might be best. On the one hand, increased competition is often seen as a key mechanism for promoting innovation. On the other hand, fragmented rail structures designed to promote competition, comprising multiple companies with different incentives and differing regulatory structures, create a potentially significant obstacle to the implementation of some of the innovations that emerge through research such as that being conducted by Shift2Rail. For example, costs may be incurred in one part of the system (say the infrastructure manager), but with the benefits felt by train operators (or the other way round). Alternatively, innovations may require up-front investment with pay-back periods that exceed the length of rail franchises or which are hard to manage within regulatory or other multi-annual funding agreements between Transport Ministries and rail infrastructure companies where funding may be constrained. More widely, the ability of rail systems to optimise from a whole life cost perspective may be impeded by the complex array of incentives existing within fragmented railway systems. This was, for example, the view formed by the McNulty review of the British rail system in 2011, which identified "fragmentation of structures and interfaces" and "ineffective or misaligned incentives" as two of the principal barriers to efficiency facing the industry (McNulty, 2011).

The purpose of this paper is to explore the incentive mechanisms that exist in the rail industry and how these might serve as barriers to or enablers of rail innovation, focused on cost reduction and/or performance improvement. Armed with a better understanding of these incentive mechanisms, our intention is then to propose changes that would enhance innovation and cost reduction appropriate for different contexts. For this purpose, we present here case studies of a wide range of examples, exploring barriers and enablers of innovation in different contexts. Britain and Sweden are explored in depth, supplemented by briefer reviews of Germany, France and Slovenia. Our case study approach involved undertaking a review of key industry reports and guidance and conducting a series of interviews with key industry stakeholders.

We begin our exploration by reviewing relevant academic and policy-relevant literature. Sections 3-7 then provide thematic reports of our five case studies, and section 8 provides our conclusions and recommendations.

2. Innovation and Industry Structure – an overview of the literature

2.1 Introduction

Innovation is, by definition, the development of new ideas, new devices and/or new methods, and in the context of rail, we are concerned about ideas, devices and methods that enable rail to achieve increased capacity, improved customer service, and reduced costs and carbon emissions (the 4 Cs). Innovation amongst Europe's railways is being spearheaded by the Shift2Rail joint undertaking. Shift2Rail is designed to guide rail research innovation into the next decade. By way of example, a number of areas of innovation were outlined by the Chair of the European Rail Research Advisory Council (ERRAC) following the launch of Shift2Rail. These include the use of mechatronics so as to move closer to a maintenance-free railway, the use of digital train control and train assured braking to enable substantial increases in capacity, the digitalisation of infrastructure and a complete rethink of train stations (Doherty, 2015).

Nevertheless, widespread concerns exist regarding the relatively slow pace of innovation in the rail industry. In establishing the Shift2Rail programme, the European Commission acknowledged problems associated with rail innovation in Europe, stemming from:

- "Fragmentation of R&I efforts;
- Low leverage of EU R&D investment;
- Limited and uncoordinated participation of stakeholders along the value chain;
- High costs, risks and lead-times of R&I investment" (EC, 2014).

Furthermore, RSSB, identify four further related difficulties:

- "A fragmented industry structure where costs and benefits frequently sit with different organisations
- A project and technology-led culture which is focused on outputs rather than outcomes
- High aversion to risk, making the approval process for new products and projects complicated and long
- Limited resources for testing and developing new ideas" (RSSB, 2017).

In the face of these concerns relating to fragmentation, coordination and cultures, we have sought to review some of the key literature in these areas as they relate to incentives for innovation.

2.2 Industry Structure and Incentives in Principle

There is ongoing debate regarding the pros and cons of vertical separation versus integration in a number of industries in which part of that industry's activity is competitive and other parts exhibit monopoly power. The concern is that an integrated monopoly will exercise its market power to restrict supply and raise prices or that it will, in short, behave anti-competitively. Typically, this concern is expressed in relation to network industries in which the network infrastructure exhibits natural monopoly, whilst services on that infrastructure may be provided competitively. Examples include telecommunications, electricity and gas, as well as airlines and railways.

The principal arguments on either side are well-rehearsed. In favour of separation, it is argued that it enables competition and, thereby, provides incentives for cost reduction and innovation. Moreover, many argue that it is only with a wholly vertically separated structure that non-discriminatory access to the infrastructure can be assured, thereby enabling that cost-reducing and innovation-stimulating competition to occur. On the other hand, the principal argument in favour of integration is that it promotes system optimisation (Pitman, 2007). The key ways in which it does this are identified as being by reducing transactions costs (Merkert et al 2012) and by removing misalignment of incentives (McNulty, 2011). Caves and Doyle (2007) distinguish between functional and structural separation, though they find that, in terms of the coordination of investment, the two forms of separation appear to give rise to similar or identical challenges

In their review of theoretical literature, Caves and Doyle (2007) identify the potential problem of opportunistic behaviours. First highlighted by Klein, Crawford and Alchian (1978), the problem arises where one party waits for a partner organisation to purchase a specialist asset, and then seek to reduce the price paid for the output of that asset, in the knowledge that their partner has no alternative outlet for this output. Klein, Crawford and Alchian (1978) showed that the scope for such behaviour increases as assets become more specific. Caves and Doyle (2007) do, however, point out that whilst this could be solved by vertical integration, it could also be solved via contractual arrangements within a vertically separated structure.

Much of the theoretical literature points at the difficulties and costs of writing and enforcing these contracts, and of associated regulatory arrangements, and suggests that investment coordination and market power problems will persist even under vertical separation. However, Caves and Doyle (2007) label these theoretical assertions as 'contentious', and propose that they should be tested in practice, within the specific contexts of industrial sectors. They also argue that "Creating incentives for efficient investment is a major problem under existing systems of sectoral regulation.. [and that] …. These well-known difficulties are independent of the vertical structure of the industry and should not be attributed to it" (Caves and Doyle, 2007).

OECD (2001) has recommended that the appraisal of alternative structures for a firm be based on their relative merits, noting that separation has a big benefit of eliminating anti-competitive practices. However, in its review of different forms of separation (ownership, cf Club Ownership cf Operational Separation), it notes that one disadvantage of Operational Separation is that the 'possible lack of profit motive reduces incentive to provide innovative and dynamic services" (OECD, 2001).

2.3 Industry Structure and Incentives in Other Industries

In the context of investment in innovation in telecommunications, OECD (2003) identified concerns regarding the incentives for adequate investment in network infrastructure under vertical separation. This concern appears to stem from the idea that separation interrupts the flow of revenues generated by the investment, preventing them in whole or in part from accruing to the infrastructure provider. Separation is thereby seen as inserting a requirement to coordinate investment amongst the wholesale and retail operators, serving as an impedance to investment in innovation. They assert that in telecommunications, an industry in which the pace of technological change and demand growth are rapid, such investment coordination problems could be considerable. OECD's judgement appears to be that whilst separation is likely to promote competition which 'could' in turn promote network enhancement, this would come at the cost of an erosion of incentives to upgrade the network" and have a "negative impact on broadband development" (OECD, 2003). Furthermore, they highlight that industry observers have suggested that in many countries, stock market sentiment may mean that "it is the incumbent with a steady cash flow that could be in the best financial position to enhance the local network" (OECD, 2003).

Caves and Doyle (2007) argue that whilst evidence from the regulated sectors is mixed, evidence from the non-regulated industries, including the personal computer and gaming industry, is that contracting arrangements can be used effectively to manage vertical relationships;

In summary, Caves and Doyle's review leads them to conclude that "by definition, a separated structure imposes heavier demands on contracting, but the evidence of academic research cited and case studies presented is that contracting can, in most cases, take the strain, by such means as long term or risk sharing contracts" (Caves and Doyle, 2007).

2.4 Industry Structure and Incentives in the Railway Industry

For railways, recent evidence in relation to this debate suggests it depends on traffic density, with more densely trafficked railways benefiting from integration and less-densely trafficked railways benefiting from separation (Mizutani et al, 2015)

Nash et al (2014) utilise a value chain model which identifies 4 distinct planning terms and coordination circles;

- Long Term Planning, with a focus on strategic investment;
- Medium term service planning,
- Short term timetabling; and
- Real time adjustments.

In vertically separated industries, there is scope for misalignment of incentives to occur at each of these phases. That is, players have an incentive to optimise their own costs rather than the costs of the system as a whole. For our purposes, misalignments in the longer and medium term phases are probably most relevant to the issues of innovation. These include misalignments in relation to upgrading or downgrading infrastructure (phase 1), to the quality of resources used, with knock-on impacts on performance (in phase 2)

Nash et al's review indicates a number of possible consequences of misalignments, including: "held-up investment opportunities in various technical assets, networks not developed in line with market requirements and suboptimal combinations of assets (rolling stock, track and personnel)" (Nash et al, 2014). In turn, these consequences lead to inflated production costs and investment externalities, though quantitative evidence on the actual cost consequences is very limited.

Acknowledging these misalignments flowing from vertical separation, Nash et al identify a number of mechanisms put in place across Europe, combining market and hierarchy. For instance, track access charges are an attempt to re-align incentives but "it appears impossible to design a track access charging system that simultaneously provides for non-discrimination, appropriate incentives for efficient development of the network and appropriate incentives for its use" (Nash et al, 2014). In addition to track access charges, there are also performance regimes, long term contracts, strategic partnerships and joint ventures, though these come with transaction costs.

Hence, whilst track access charges and performance regimes can play a part in re-aligning incentives, it does not appear that they can do so entirely, and so a range of other initiatives to incentivise innovation is therefore being put in place.

2.5 Key Messages

Across many industries, there are similar concerns about how to balance the benefits and costs of vertical separation, and these are particularly acute in regulated industries involving networked infrastructures and an element of natural monopoly. Generally, concerns have not been expressed explicitly in relation to innovation, but we have interpreted broadly-expressed concerns about investment as encompassing investment in innovation. Linked to this, there are ongoing debates regarding the ability of appropriate contracting arrangements within vertically separated structures to allow the competitive benefits of separation to be realised, whilst securing the coordination benefits usually associated with integration. Beyond this, there are also concerns that problems are being assigned to structure, when they may actually be about regulation, and perhaps adequacy of contracts.

Whilst it is important to engage with the theoretical context and interesting to examine experience from other industries, it becomes clear that the outcome of the debates in relation to these issues is closely associated with the specifics of each industry. Also, as mentioned above, there is clear scope for further exploration of the specific issues as they relate to innovation, as compared with investment more broadly. Hence, in the subsequent chapters we seek to complement the econometric work of Mizutani et al (2015) and build on the more qualitative work of Nash et al (2014) in order to gain a much more detailed understanding of the specifics affecting the rail industry in a range of different settings, and to closely focus attention on the issue of incentives for innovation.

3. Incentives for Rail Innovation in Britain

3.1 Introduction

The structure of the rail industry in Britain is as follows. Network Rail, now firmly classified as being in the public sector, is the infrastructure manager. Passenger services are almost entirely provided under franchises, typically for seven years. Freight services are provided by commercial operators under an open access regime. Most rolling stock is leased from separate leasing companies or increasingly from manufacturers. The Office of Rail and Road is the regulator, with a particular remit to approve track access charges and to assess the revenue requirements of Network Rail on the assumption that it operates efficiently. This it does through a five yearly periodic review (PR) which sets out what Network Rail is to achieve in the coming five year control period (CP) and what funding it will get to do it, via a combination of track access charges and network grant from the government, in the light of the government's high level output specification (HLOS) and statement of funds available (SOFA). In the past, this process covered both renewals and enhancements, but for the coming control period enhancements will be considered separately as and when they are fully costed and the business case established. Finally, the Rail Safety and Standards Board is an industry body which advises its members not just regarding Safety and Standards but more generally on issues spanning the railway as a whole, including specifically for promoting innovation.

Ongoing concerns about the industry's structure and the effectiveness of the industrial processes have led to a number of reviews being conducted in recent years, including those of McNulty (2011), Brown (2013) and Shaw (2015). Amongst the changes these have led to are the setting up of the Rail Delivery Group to bring together Network Rail and the train operators in planning, the introduction of alliances between the infrastructure manager and individual train operators to try to improve efficiency and – now underway – the allocation of much more responsibility to the individual route management within Network Rail, whilst retaining a central systems operator responsible for charging and allocation of capacity,

Our case study of Britain involved reviewing key industry reports and websites and conducting interviews with representatives from each of the key strategic industrial stakeholders. From our review of documentation, we have observed three broad sets of incentives of possible relevance for rail innovation. Firstly, there is what ORR refers to, in its Final Determination for PR13, as a 'package of incentives', comprising charges, financial incentives and contractual incentives, the stated purpose of which is to promote efficient behaviour across the industry. These include track access charges, performance regimes and regulatory outputs required from Network Rail.

Secondly, beyond this incentive framework associated with the regulatory Control Period, there are also a number of funding initiatives which can also serve as a spur to rail innovation. These include:

- National stations improvement programme;
- East Coast Connectivity Fund;
- CP6 Development Fund;
- Network Rail Discretionary Fund for Small Local Schemes;
- Strategic Freight Network Fund;
- Passenger Journey Improvement Fund;
- Innovation Fund;
- Strategic Research and Development Fund; and
- ETCS Cab Fitment Fund.

Thirdly, it has been observed by many of our interviewees that reputational incentives may also play an important role. Certainly reputation is important to career advancement, whilst the fact that managers have to face strong criticism from the media and from bodies such as the Public Accounts and Transport Committees of the House of Commons is undoubtedly an incentive. Indeed, reputational incentives form an important element in the decentralisation within Network Rail, to permit yardstick competition between the routes.

Crossing over with some of the items above, our interviewees identified four routes they felt to be particularly important, by which investment in rail innovation can be pursued:

1. Investments committed at the time of the PR and added to the Regulatory Asset Base (RAB)
2. Further investments added by agreement with DfT and added to the RAB.
3. Investments proposed by TOCs which are ultimately paid for by them through additions to track access charges
4. Specific additional funding mechanisms, such as Innovation in Franchising, Spend to Save, and the specific £50m innovations fund included in the previous High Level Output Statement (HLOS).

However, following on from the reclassification of Network Rail, these mechanisms have changed significantly. In particular, the ability of Network Rail to borrow from the market to finance schemes added to the RAB has been removed. Network Rail can now only borrow from government and a strict borrowing limit has been imposed.

In the following paragraphs, we comment in more detail on these mechanisms.

3.2 Track Access Charges

In Britain, track access charges are determined by the Office for Rail and Road (ORR), as part of a periodic review process. This process determines the outputs that the infrastructure manager – Network Rail – must deliver, the level and make-up of income that Network Rail can collect in order to fund its activities, be that from charges or other sources, and a number of mechanisms designed to incentivise Network Rail to deliver and 'out-perform' its commitments. These 'Regulatory Outputs', 'Revenue Requirement' and incentive mechanisms are set for a 5 year 'Control Period.

The current set of track access charges have a variable and a fixed component, and are predominantly cost-based. The variable component comprises a number of elements designed to recover the marginal costs of wear and tear to the infrastructure and the cost of congestion (the increased incidence of reactionary delay as a result of an additional train on the network). There is also a mark-up on certain freight commodities designed to contribute to the avoidable fixed costs of freight traffic. Fixed costs are only paid by passenger franchisees and are based on an allocation of part of the fixed costs of the system.

For Network Rail, an innovation which reduces marginal cost (including congestion cost) will benefit them for the duration of the control period, but at the next periodic review is likely to be passed on to users in the form of a reduced charge. Thus the incentive provided by this system is very much for short run cost reductions rather than longer run innovations. To the extent that such innovations involve capital cost, Network Rail should be recompensed for that through the overall settlement, and indeed the costs of wear and tear are assessed as life cycle costs of a 35 year period, but at the present time – with Network Rail up against its spending limits – it may be unaffordable. An example of such an innovation is being considered in the 'Steels' project, in the form of a new type of steel with higher first cost but lower maintenance cost. However, given that franchised train operators do not benefit from such cost reductions (or lose from cost increases) as these are allowed for in adjustments to payments under the franchise agreement, they have no incentive to help Network Rail secure such cost savings, Incentives are greater for freight operators and open access passenger operators (i.e. those operating without a franchise) as they do bear any changes in track access charges themselves.

As noted above, arrangements may be agreed for improvements in the capability of the infrastructure in terms of speed, capacity or performance that are paid for directly by the train operator as increased access charges. Such arrangements should place appropriate incentives to innovation on both parties. However, within a franchising system with relatively short franchises, any arrangement for an operator to contribute would need to be established so that it was distributed over time and flowed through from being an arrangement with an incumbent franchisee to being an arrangement with any subsequent franchisees within the payback period of the investment. Otherwise, the possibility of subsequent franchisees 'free-riding' may serve as a disincentive to the incumbent franchisee making that contribution. Similarly, other train operators may benefit from the scheme, leading to an incentive to all operators to try to free-ride rather than pay.

Indeed, our interviewees highlighted cases where investments have been proposed by train operators and ultimately paid for by them through additions to track access charges. Notable cases include Chiltern Evergreen [which however took place under a much longer 20 year franchise] and, on a smaller scale, a programme of multi storey car parks on the West Coast Main Line amounting to an investment of approx. £100m [again, as part of the terms of a longer franchise period]. Our interviewees suggested that these sorts of projects can extend beyond the length of the franchise if DfT commits the future franchisee to pay via a residual value mechanism, although we understand that in practice this has proved problematic.

Under a regime in which track access charges are based on marginal cost, there is no incentive on the infrastructure manager to attract more traffic (and indeed maybe a considerable disincentive if it makes achievement of performance targets more difficult). Mark ups may provide such an incentive (at least in the short run until the next periodic review). Britain also has an explicit incentive payment based on traffic volumes, but interviewees thought that this was too small to have much impact.

In principle, track access charges, together with their associated incentive mechanisms, provide price signals to train operators and policy makers regarding the efficient use of the infrastructure and, as such, affect the ways in which operators and infrastructure manager work together. The high degree of differentiation of charges according to rolling stock characteristics should provide incentives for train operators to invest in track friendly rolling stock wherever the additional cost is offset by the saving in wear and tear cost, although short franchises may limit this incentive. A previous study (Nash et al, 2014) found that franchised train operators had short time horizons and were more concerned about proven reliability than life cycle costs. Recent public sector led procurements – IEP, Thameslink and Crossrail – have explicitly used life cycle costs as a criterion, although these have been controversial for other reasons and doubt has been expressed as whether the most cost-effective solution has always been found. Presumably the problem of short time horizons is less acute with freight and open access operators, whose time horizons are not shortened by franchise expiry dates.

However, franchisees have limited scope to respond to price signals given by the track access charges, as they are bound by their franchise agreements and, hence, often have very limited flexibility regarding what services they offer, and how they offer them. Moreover, currently franchisees are fully compensated for any changes in charges during the franchise, so they have no incentive to cooperate with Network Rail to reduce infrastructure costs, for instance by adjusting timetables to give longer possessions. For this reason, RDG recommended that the next review of charges and incentives deliver a charging and incentive regime that aligns better with other industry arrangements – most notably perhaps, the arrangements of the franchising system. Also, ORR has suggested that Train Operating Companies should be exposed to infrastructure cost risk. For this to be fully effective, however, would require a charging system which related fixed charges to costs the train operating company could influence whilst fully charging all costs. In the presence of joint costs, this is not possible, and the presence of joint costs creates possibilities for one train operator to free-ride on the actions of another.

In short, because track access charges are calculated to be cost-reflective, and given that Network Rail is a monopolist, there is a danger that Network Rail’s long run incentives to increase productivity, control costs and innovate are limited. Furthermore, because the variable component of the charge does not cover all of the costs of providing capacity, Network Rail’s incentive to act commercially in response to increased demand for the use of its infrastructure also risks being limited. Recognising these limitations, ORR puts in place supplementary financial and contractual incentives, which we discuss in the next section. It is also engaged in a review of track access charges which may lead to a substantial change of approach in the next control period.

3.3 Financial and Contractual Incentives

A number of related financial incentives exist alongside track access charges, aimed at promoting efficient behaviour across the industry. ORR set these out in their PR13 Final Determination as follows:

“(a) developing the existing efficiency benefit sharing mechanism into a route-level efficiency benefit sharing (REBS) mechanism. This incentive is designed to strengthen the alignment of incentives between Network Rail and train operators – through the development of a default mechanism in CP5 for Network Rail to share efficiencies with train operators – in order to support greater co-operation to drive down industry costs. It works by allowing efficiency gains or losses to be shared between Network Rail and its customers (i.e. operators) on an annual basis;

(b) asking franchising authorities to provide new franchises with exposure to technical or cost-reflective (as opposed to policy related) changes in the variable usage charge at future periodic reviews. Also, working with governments to explore how we can increase franchised train operators‟

exposure to the fixed charge and to changes in it. The rationale is similar to that for REBS but the mechanism works by giving operators a greater interest in infrastructure costs at a periodic review;

(c) strengthening the incentives for the industry to work together to drive down the costs of enhancements and to align scope, specification and delivery of projects better with the needs of the operational railway and its customers. ORR wants Network Rail and operators to enter into commercial agreements that will help Network Rail to achieve improvements and reward both parties if these are achieved;

(d) supporting investment in R&D and innovation by introducing a matched-funding financial incentive; and

(e) developing the existing volume incentive mechanism in terms of both its design and payment rates in order to improve its effectiveness. The volume incentive is designed to encourage Network Rail to consider unexpected demand from its customers and in doing so to make trade-offs similar to those made by a company operating in a more commercial setting." (ORR, 2013)

Our interviewees commented specifically on the REBS and the Volume Incentive. In general, they felt the REBS to be not very effective. It is understood that train operators dislike the mechanism because they have to pay if Network Rail's costs rise above the baseline, even if they – the train operator - have no way of influencing them. Whilst train operator liability under the scheme is capped, Network Rail costs are currently so far above the baseline that train operator contribution is, in every case, at the cap. In this situation, there is no impact on train operators as a result of marginal changes in costs. Consequently, any incentive effect is largely removed. Therefore, it is anticipated that the REBS will be reduced in scope to make them more specific and achievable. For instance, it might be replaced by a scheme relating specifically to expenditure that a train operator can influence (e.g. renewals). In this case, however, any incentive will impact on a lower proportion of expenditure.

Regarding the Volume incentive, whilst this is designed to incentivise small projects to increase capacity, interviewees highlighted that currently Network Rail is more concerned about the impact of additional trains on performance than possible revenue from the volume incentive. This is despite the Incentive having been 'strengthened' as part of PR13.

The Possessions Regime Schedule 4) and the Performance Regime (Schedule 8), represent contractual incentives designed, in the words of ORR, to:

"(a) compensate train operators for the financial impact of planned and unplanned service disruption attributable to Network Rail and other train operators;

(b) help align incentives between Network Rail and train operators, so the impact of service disruption on revenue and/ or costs is incurred by the organisation who cause the disruption, rather than the train operator that faces the disruption; and

(c) provide appropriate signals so as to drive the decision-making in relation to performance and possession management, for example, in relation to where to make investments, or to give an indication to Network Rail on whether it is better to have a short possession but with higher engineering costs or take a longer possession" (ORR, 2013).

Our interviewees commented favourably on the ways in which these contractual incentives are functioning, acknowledging them both as essential protection from risks outside their control (and noting that the premium paid to or subsidy paid by DfT under the franchise agreement does not change according to how well Network Rail performs) and as an important mechanism for aligning industry incentives and potentially, thereby, fostering innovation. It was noted that Schedule 4 is particularly effective at encouraging advance planning by Network Rail, which helps train operators plan ahead (there is a lower penalty for possessions planned well ahead than if possessions take place at short notice). Also, it was felt that Schedule 4 serves to drive the way in which renewals are done, for example where high output ballast machines are employed on the network. On the other hand, it was noted that these arrangements only give Network Rail an incentive to consider loss of revenue to the operator and not wider economic and social implications of poor performance, and may not be large enough to be effective in all cases. This is a particular issue for commuter services, where elasticities and therefore loss of revenue, are low but political pressure for improved performance strong.

However, importantly, these incentives are long run; Network Rail would expect them to continue indefinitely. But, it was also recognised that the strength of these incentives, and Network Rail's ability to respond to them, is constrained by their ability to fund maintenance and renewals activities, and any investment initiatives that might impact on maintenance and renewals expenditures. It was noted, for instance, that the current cash constraint may well have focused the minds of Network Rail on whether they have the right balance of maintenance and renewals – given that the optimal balance between the two is not totally precise. Indeed, it is understood that, in the face of current financial pressures, renewals are being deferred and it is recognised that, as a consequence, performance will suffer (with train operators being compensated for this via Schedule 8).

More generally, our interviewees identified that Network Rail is currently not achieving its performance targets and may be penalised by ORR for this. In this way, the need for Network Rail to achieve its Regulatory Outputs would seem to be a strong incentive, not only from the point of view of these contractual incentives but also from the point of view of management bonuses and managerial reputation. It is also relevant here to note that, given that these incentive mechanisms ensure that it is in both the interest of train operators and Network Rail to work together on performance issues, there are joint task forces on improving performance.

3.4 Innovation in Franchising

It is recognised that, whilst the franchise competition itself may give strong incentives for innovations to come up with a winning bid (particularly since the award of the franchises now takes account of quality as well as financial results) a relatively short franchise life gives little incentive to innovate after the franchise is awarded.

The Innovation in Franchising initiative is a pilot scheme aimed at encouraging franchisees to invest in innovation during the life of the franchise by means of a requirement for franchisees to include an innovation strategy in their franchise bids (and giving it greater weight in the evaluation of bids).

However, within the period of the franchise, operators are naturally geared to delivering the terms of the franchise agreement and nothing more. Moreover, the profit sharing arrangements incorporated in many franchises limit the scope to make additional profits even in the short run. The residual value provision does help give longer run incentives for physical assets such as car parks, but it would seem much more difficult to apply in the case of softer factors such as marketing or improving the customer experience. Leasing arrangements also may make it difficult to modify trains without paying high fees (e.g. IEP).

The innovation in franchising programme is designed to counter this, and involves earmarking 1% of franchise revenue for spending exclusively on innovation projects. The fund is administered by RSSB's Innovation programme, and relies on the franchisee submitting proposals to RSSB, for scrutiny by an Innovation Board, comprised of senior industry figures. The pilot programme involves Virgin East Coast, Northern and Transpennine and totals £50m over 3 years. After its pilot phase the programme will be reviewed before deciding whether to continue as part of future franchises.

3.5 Ring-Fenced Funds

A number of funds have been put in place to provide for infrastructure enhancement initiatives, including:

- National stations improvement programme fund – established in 2009 with £150m government funding, to co-fund projects at stations to improve customer experience, Network Rail has ring-fenced £70m to continue this programme during CP5;
- East Coast Connectivity Fund;
- CP6 Development Fund;
- Network Rail Discretionary Fund for Small Local Schemes;
- Strategic Freight Network Fund – further details given below;
- Passenger Journey Improvement Fund;
- Innovation Fund – there was a specific item in the DfT's High Level Output Specification (HLOS) to fund innovations that "drive cost reduction and quality improvements" (ORR, 2013) although following the Hendy review into the financial difficulties with Network Rail's investment programme this was abolished;
- Strategic Research and Development Fund – where Network Rail uses funds from third parties or outperformance to invest in R&D and innovation, a commitment was made in the Final Determination (ORR, 2013) that DfT would provide matched funding of up to £50m; and
- ETCS Cab Fitment Fund – again, there was a specific item in the DfT's HLOS for Network Rail to fund train operators to install ERTMS in cabs.

Our interviewees highlighted the 'spend to save' scheme as being a potentially useful initiative in relation to rail innovation. The scheme is for projects identified after the start of the control period. Despite its title, the scheme can include revenue generating projects. However, with a value of approximately £30m, it is relatively small. Our interviewees highlighted that one initiative funded via this scheme was the Mountfield project, for Network Rail to acquire and lease out freight yards and terminals to make access easier than when owned by individual operators (usually DB Schenker) who were not using them. (Network Rail can also use this scheme to fund train operators to do things which will reduce NR costs, for example to put cameras on trains to inspect overhead line).

3.6 Innovation in Rail Freight

In its 2016 Rail Freight Strategy, the Department for Transport set out four key priorities for government, the industry and others to take action on in order to maximise rail freight's potential. These priorities are:

- Innovation and skills;

- Network capacity;
- Track access charges; and
- Telling the story of rail freight.

With regard to innovation, the need for the industry to innovate in relation to the markets it serves and the business models it employs, as well as developing "new technological solutions to improve efficiency and tackle wider challenges for the industry" (DfT, 2016) are highlighted. As an example of innovation in relation to markets and business models, two interesting cases are described: firstly, the case of Colas Rail Express Deliveries, which involves trialling the use of converted rolling stock for the carriage of supermarket produce and cars from central warehousing direct to Euston Station; and secondly, the case of DB Cargo and CEMEX's 'Pop-up' Rail Depot, installed adjacent to the West Coast Main Line in a matter of weeks using a ready-made weighbridge and office. However, these cases seem to have taken place independent of any specific incentive mechanism.

In relation to facilitating technological development, the cross industry Freight Technologies Group was established in 2014. This group includes Network Rail and the freight operators and has a remit to develop schemes that improve performance, safety and the customer experience, reduce costs and enhance capacity. At present, the Group is delivering three such schemes, involving timetable advisory software, Freight Collaborative Decision-Making and the Mobile Consisting Application, though again, this does not appear to be as a consequence of any specific incentive mechanism.

One further development regarding innovation in rail freight, highlighted in the DfT Strategy is the Data for Freight project. This is a collaborative effort between Transport Systems Catapult and the DfT, aimed at making use of data science techniques and closer working between freight operators to make better use of data and collect better data so as to facilitate improvements in efficiency and resilience.

In order to encourage investments to enhance infrastructure for freight, the Strategic Freight Network Fund was established in 2009. During CP5, this was allocated £253m, though this was reduced to £235.9m as part of a Network Rail review in January 2017. Projects funded through this fund have included major chord connections at Nuneaton North, North Doncaster and Ipswich, gauge clearance on the Strategic Freight Network to allow for bigger containers to be carried on standard wagons, expansion of the capacity of the branch line to the Port of Felixstowe and improving rail access to the Port of Liverpool. There are some concerns, expressed in DfT's Rail Freight Strategy, regarding the incentives linked to these sorts of projects. In essence, there is a question of whether there needs to be some safeguarding of capacity for freight, where these sorts of freight-related enhancements have been put in place (or where forwarders have invested in warehousing or depot facilities adjacent to the rail network), in order to incentivise the rail freight sector to buy into these schemes and make best use of them. In any case though, these sorts of projects would appear to be more straight-forward investments in infrastructure capacity, rather than cases of rail innovation.

3.7 Alliances and Devolution

McNulty concluded that one major reason for the growth of costs in the early years of this century was a misalignment of incentives between train operator and infrastructure manager, with neither incentivised to do what was best for the rail system as a whole. This is clearly relevant also to the issue of innovation. Arguably one of the barriers to innovation is that very often the costs and benefits fall on different organisations within the industry.

Two of the recommendations of the McNulty report were more devolution of control of regional services and closer links (alliances or joint ventures) between franchisees and the appropriate parts of Network Rail.

Devolution of control is spreading, with franchises let by the Scottish government, Transport for London, Merseyside PTE, and a partnership of Department for Transport and Rail North (a coalition of 29 local authorities in the North of England) and similar changes in Wales and the West Midlands to follow. Of these, we only conducted interviews in Scotland. There we gained the impression that devolution had been very beneficial, leading to a degree of involvement by the franchising body in long term planning for the rail system as a whole which was difficult to achieve in the days when twenty five franchises covering the whole system were all let in London, and which should make innovation easier. Of course the fact that infrastructure is also devolved in Scotland also facilitates planning for the railway as a whole and permits the sort of total route modernisation which is very helpful in implementing innovations which span the wheel-rail interface.

Although many alliances have been signed between train operators and Network Rail on specific issues, only one to date has been the sort of ‘deep alliance’ that involved sharing cost and revenue differences from an agreed baseline, and therefore completely removing the issue of incidence of costs and benefits (although only for the lifetime of the alliance). That is the Wessex Alliance between South West Trains and the appropriate Network Rail route, and may be regarded as one of the most innovative developments organisationally for many years. In this case, the staff of the two organisations, whilst remaining employed by either South West Trains or Network Rail, was effectively merged into a single structure under a single Managing Director.

Although it was argued that there were some cost savings as a result, the primary aim of the alliance was to tackle the deterioration of the infrastructure which was causing declining performance. Thus an early decision was to resume ballast cleaning and tree clearance, which had been abandoned by the infrastructure manager many years earlier. Another benefit was the ability to plan maintenance work and renewals in the way that was most efficient for the railway as a whole, which was in practice to rearrange services to provide longer possessions. (Although renewals were not formally part of the alliance, the fact that there was a common Managing Director and chain of command facilitated this).

However, there were clear limits to the ability to innovate imposed by the fact that Network Rail as an organisation applied national standards rather than permitting what was most appropriate for the circumstances (for instance, a densely used commuter railway with third rail electrification is very different when it comes to track maintenance and renewals from an overhead electrification main line, and different solutions will be best). As noted above, the alliance was also limited in its ability to take long term decisions by its short life. It was intended to last until the end of the franchise five years after it was signed, but was in fact terminated after only three years. A long term vertically integrated concession in which the infrastructure was leased to the operator would overcome these problems, although raising other issues such as ensuring fair treatment of other operators over the track, particularly where they constituted a substantial part of the traffic.

The premature dissolution of the Wessex Alliance (whilst leaving some of its successful features such as joint control centres and planning in place) may be taken as illustration of the difficulties of reaching agreement to implement such a structure between a private sector train operator (which may be sceptical about the ability of Network Rail to achieve planned cost savings) and a publicly owned infrastructure manager (which may be sceptical about the train operator's demand and revenue forecasts.

It was a condition of the award of the ScotRail franchise that an alliance be sought with Network Rail, but agreement could not be reached on a deep alliance sharing costs and revenues. Moreover the wish to act quickly precluded the sort of complete reorganisation of staff into a single structure that had occurred in the Wessex Alliance. However, a single managing director was appointed, becoming a member of Network Rail staff (so as to have the influence within that organisation such a senior position brought with it) but also managing the ScotRail subsidiary of Abellio, also with single heads of performance and train control. It was argued that a single Managing Director could bring together the team of senior managers of both organisations on a regular basis and that ways could be found of ensuring that solutions that were good for the railway as a whole produced outcomes which adequately rewarded both partners.

3.8 Other initiatives

Our interviewees highlighted two specific funding programmes designed to encourage innovation on rail, one run by Innovation UK on behalf of the Department For Transport (DfT), entitled "Accelerating innovation in rail", and one by RSSB, entitled "Enabling Innovation". The Accelerating innovation in rail programme comprises specific calls designed to further the achievement of the railway technical strategy. The most recent call, published in March 2017, invited consortia to bid for a share of £9m to deliver "high-value, low-cost rail solutions or improve user experience through stations" (Innovate UK, 2017). The intention is that 60% of projects should be local (i.e. a specific TOC) and 40% national (i.e. a wider consortium) but this is apparently not strictly adhered to. The RSSB programme states that it covers "innovation beyond technology and R&D, to include the people, culture and processes involved" (RSSB, 2017). The RSSB programme was formerly jointly funded by both Network Rail and DfT but this is now solely DfT. However, the scale of resource for these initiatives has been criticised. For instance, the £9m available via Innovate UK represents less than 1% of industry turnover, as compared with funding amounting to approx. 5% for some of the world's main rolling stock manufacturers (Jack, 2017; McNulty, 2011). It has also been argued that the administrative mechanisms and the fact that research is often done by those remote from the actual practical problem hamper effectiveness (Hacktrain BARRIERS report). For the future, there are proposals to develop a network of centres of excellence relating to different areas of railway research and innovation, under the banner of the UK Rail Research and Innovation Network, and industry partners are said to have committed £62m to this initiative.

3.9 Conclusion

The reform of the British rail network paid more attention to seeking to give the correct incentives to each part of the industry than was the case in most other countries. Thus Britain implemented a complex track access charging system, sophisticated performance regimes and gave the regulator the power and responsibility to investigate the efficiency of the infrastructure manager in order to determine a tough but achievable package of regulatory outputs and funding.

Nevertheless, there are three essential problems with the incentives for innovation in the British rail industry. Firstly is its fragmentation, so that the balance of costs and benefits from innovation may fall differently on different partners. This effect is amplified by the failure to expose train operators to infrastructure cost risk, although the presence of joint fixed costs means that in any event it would be difficult to reflect fully the costs under the influence of each train operator in the charges (fixed and variable) they pay. Moreover, adding infrastructure cost risk to those borne by train operators might reduce the level of competition for franchises and on the track. Alliances were seen as the solution to the problem of misaligned incentives, but they have proved difficult to negotiate and maintain (However, there might be more incentive on train operating companies to form alliances if they did bear more infrastructure risk). Secondly, the fact that infrastructure and operations are planned separately and to separate timescales makes taking a total railway approach to the modernisation of particular routes difficult, although this may be needed if innovations span the track-train interface. Thirdly is the problem of short time horizons. Both the regulatory regime, with its emphasis on five year control periods, and the franchising system, with typically seven year franchises, emphasise short term results rather than long. An added problem is that control periods and franchise lengths do not coincide, leading to additional lags in any impact measures introduced for one control period might have on franchises.

It has been suggested that a solution to all these problems may lie in long vertically integrated franchises, in which the train operator leases the infrastructure from Network Rail. But this raises problems of its own (dealing with risk in long franchises, ensuring appropriate long run stewardship of the assets, particularly towards the end of the franchise and protecting the position of other operators over the system) and the balance of advantages will differ according to circumstances (particularly how far the services over the section of infrastructure can be appropriately concentrated into a single operator).

In the meantime, there is a particular problem caused by the fact that Network Rail is right up against its financing limits and therefore cannot afford any innovation that requires capital expenditure. Thus currently the only route for funding new innovations in infrastructure is by third parties. (E.g. new stations funded by local authorities). The constraint on levels of borrowing is one driving force behind the interest in concessioning individual routes as a way of bringing in private sector funding (but such schemes still seem a long way off except for East West Rail, where the Department of Transport is looking for a private funder). Leasing options, such as a steel manufacture funding renewals and leasing their use to NR, are also possible ways round the borrowing constraint.

It has been put to us that tight funding constraints are themselves a massive incentive for efficiency and innovation. However, it is to be hoped that however tight the funding constraints are for the next control period, room can be found within them for some earmarked funds to promote innovation and to permit sensible investment where this will reduce life cycle costs.

3.10 List of interviewees

We are very grateful to the following for agreeing to be interviewed; we learned a great deal from them, but responsibility for the report is solely our own.

- Bill Davidson (Rail Delivery Group)
- Richard Davies (Department for Transport)
- Howard Farbrother and Martin Brennan (Rail Safety and Standards Board)
- Chris Hemsley and Emily Bulman (Office of Rail and Road)
- Alex Hynes (ScotRail Alliance)

4. Sweden

4.1 Organization of Sweden's railway industry

Four different actors on the scene of Sweden's railway industry are relevant for this project. The first is the firms that provide train services, the operators. All freight services are provided on a commercial basis while the provision of passenger services is split 50/50 between commercial and tendered traffic. This inter alia means that operators pay track user charges, they compete with freight and subsidized passenger operators (see below) for access to tracks in parts of the network with high capacity usage and they purchase their own rolling stock.

Train services are operated on a railway infrastructure supplied by *Trafikverket* (Swedish National Traffic Administration, the IM). This is done on behalf of the government, as further specified in section 4.2. Section 4.3 specifies the assignment given to a third player in the industry, namely *Transportstyrelsen*, Sweden's regulator. The government offices, as represented by the Ministry of Enterprise and Innovation, is the principal of both the IM and the regulator; this will subsequently be referred to as the government.[4] *Trafikanalys* (Transport Analysis) is another independent agency and is working in a very close relationship with the government (section 4.4).

The SERA Directive establishes the framework for the way in which member states are supposed to organize their railway industries, and *Järnvägslagen* (2004:519) is the Act that codifies the core parts of the Directive for Sweden. The Act establishes principles while the government is responsible for establishing the organisation that is necessary for implementing it. This is done by way of instructions to the country's agencies; these instructions are specified below. All tasks defined by the Directive and in the Act but that are not delegated to agencies are by default the responsibility of the government.

Section 4.5 summarizes some findings of a committee that has considered the efficiency of the IM. Section 4.6 provides a description of the way in which Sweden has designed a performance scheme. This is a mandatory SERA instrument for enhancing quality in train service provision. Another mandatory means is the need for member states to use a multi-annual contract or other regulatory arrangement to promote efficiency; this is considered in section 4.7. Section 4.8 summarizes the institutional structure of railway service provision in Sweden.

Section 4.9 presents one organisational change that has had consequences for the overall incentives issue in Sweden, namely the transfer from maintaining the infrastructure using in-house resources to competitive procurement that started in 2002. This is one feature that makes Sweden's way to deliver railway infrastructure services different from that in most other member states. Our take on this change concerns its consequences for innovation incentives and the incentives of commercial enterprises to innovation within the framework of them working with five-year contracts. Section 4.10 elaborates on examples of innovations that may pass the BC test. The examples deepen the brief description of examples given in section 1.1. Section 4.11 contains a brief concluding comment.

[4] For more detail about the way in which Sweden's public sector is organised, see http://www.regeringen.se/other-languages/english---how-sweden-is-governed/

4.2 The Infrastructure Manager

Sweden's IM, *Trafikverket*, was established in 2010, integrating the previous road and railway administrations under one hat. It has two core responsibilities. The first is to handle the practicalities of long-term planning of the transport system for all modes of transport. This includes maintenance of roads and railways and investment in both modes as well as in naval fairways.

After the IM has formulated a proposal for investment and maintenance priorities for the upcoming 12 years, the tentative programme is sent on a referral process. The government subsequently establishes the program; this is repeated in four-year cycles.[5] While the program does not provide a formal guarantee for future allocations, it still provides a framework for the volume of the IM's investment and maintenance level during the next few years.

The IM's second responsibility is to implement investment and maintenance activities. Based on the priorities given by the long-term program, the parliament annually establishes the actual level of appropriations. The IM tenders the implementation of all investment and maintenance activities within this framework. The main consequence of cost overruns is that the implementation of the programs is delayed.

The IM's instruction established its formal tasks and objectives; the current version was resolved in 2010 (SFS 2010:185). With an intermodal point of departure, the IM is made responsible for investment and maintenance activities. It shall do this with the objective to create prerequisites for an efficient, competitive and sustainable transport system. This shall be based on the official transport policy framework, most recently established by the parliament in 2011.

The instruction enumerates a range of specific tasks, inter alia including the following:

- To develop, administer and apply methods for CBA analysis of infrastructure measures, including ex post reviews.
- To develop and make publicly available updated traffic and transport forecasts.
- In its capacity of client for implementation of infrastructure activities, the IM shall promote productivity, innovation and efficiency in the relevant markets.
- During the planning process, it shall apply a stepwise approach to consider the following interventions; this four-step principle is supposed to make the IM consider cheaper means for achieving its objectives before proposing costly investments:
    - Actions that affect transport demand and the choice of mode of transport.
    - Actions that make the use of existing assets more efficient.
    - Minor renewal interventions.
    - Major investment projects.
- The IM shall annually report the development of productivity in operations, maintenance and investment activities to the government.

[5] For the current planning cycle, the IM's proposal was published by the end of August 2017 and the government is supposed to establish the programme by the end of the year.

4.3 The Regulator

The instructions to *Transportstyrelsen*, Sweden’s Regulator (SFS 2008:1300) establishes its main tasks to be responsible for regulation, licensing and supervision in and of transport sector activities. In the same way as the IM, this is to be done from an intermodal perspective and to create prerequisites for an efficient, competitive and sustainable transport system. This shall be based on the official transport policy framework.

The Regulator is an amalgamation of previously separate agencies for regulation and supervision of technical and traffic safety issues in the four modes of transport. This background on safety concerns permeates the Regulator’s position relative to its assignment.

The Regulator is instructed to monitor the markets for railway services at large, including the provision of railway transport. This includes a responsibility for reviewing the efficiency of the final markets for tendered and commercial passenger services as well as freight services. The Regulator therefore publishes a bi-annual assessment of the situation in this respect. It is also given a responsibility for issues referring to “(r)equirements for infrastructure managers, traffic organizers and transport companies.” (SFS 2008:1300, 3 §, task 3)

An ongoing audit addresses the design of track user charges. The regulator reviews current charges against the formulations in the Swedish Railway Act (2005:519), which in this part is identical to corresponding text in the Directive. The regulator has therefore asked the IM for documentation of how the level of charges had been estimated as well as the IM’s plans for the level of charges the next few years. It is also asking for which registries that the IM makes use of for the design and update of the charging levels. No further information about the result of the review is currently available.

4.4 *Trafikanalys* (Transport Analysis)

Tranfikanalys is a government agency charged with providing the ministry with policy advice. Guidance is to be based on the official objectives. The agency reviews, analyses, follows up and evaluates proposed and implemented measures at the request of the Government, in effect making it a semi-independent part of the ministry.

On an annual basis, *Trafikanalys* publishes a follow-up of the transport objectives. Sweden's transport policy goals were presented in Government Bill "Goals for future travel and transport" (Government Bill 2008/09:93) and adopted by the parliament in 2009. The overall objective of the field of transport policy is to ascertain an economically efficient and sustainable supply of transport for the citizens and for trade and industry in all parts of the country. The recurrent review of outcomes relative to goals are supposed to reflect the measurable consequences of policy changes.

Appendix A reproduces a summary of the 2016 follow up of policy objectives. It illustrates that the review is performed on an aggregated level and has little relevance for the way in which the IM handles its day-to-day responsibilities in general and is even less related to the issue of cost efficiency and incentives.

4.5 The Alexandersson committee

Over a sequence of years, railway maintenance costs have increased much faster than traffic. This provides one background to the government's appointment of a committee chaired by Gunnar Alexandersson in May 2013. Its overall task was to consider the way in which the railway industry is organised against the background of cost increases and concerns over train delays. The final report was submitted in December 2015. An intermediate report, SOU 2015:42 (*Koll på anläggningen*; "Control of the facility") focussed specifically on rail infrastructure maintenance and renewal. Several observations made by this and other reports produced by the committee have direct or indirect implications for the issue of incentives in the organisation.

The committee opines that a functioning and purposeful asset register is of fundamental importance for an infrastructure manager to have good control over the condition of the plant and its use. This is the basis for calculating railway charges, for prioritization of different types of interventions and for estimating resource needs in the budget process. It is also necessary for traffic management. The committee's conclusion was that the IM does not have this comprehensive system.

Except for system build-up and support, responsibility and routines for updating the information about the standard of the plant and information management at large must have a clear organizational residence. Those who are dependent on using the information must have the responsibility and authority to decide on the entire information chain and to act with respect to system support, working methods and routines. This applies to the IM's internal routines, and in the agreements between the IM and the contractors, i.e. the entrepreneurs handling daily maintenance.

The committee emphasized the need to link a register over technical installations, delays etc. to relevant economic information. The assessment was that this is necessary to meet the requirements set out in the SERA Directive. As far as the governance framework is concerned, it is not only necessary to have a coherent structure for submitting reports but also to make independent controls or reviews of the standard of the plant. Moreover, to contribute to effective resource management, it is important to establish the causal relationship between, for example, increased or decreased reliability of different components in the system and the allocation of funds. Without information about this relationship it is difficult to assess for instance whether punctuality could be achieved with more limited resources.

The government has acted on a limited number of issues raised by the committee. One example concerns the competitive tendering of railway maintenance (described in section 4.9) and the fact that the IM does not have staff of its own for performing infrastructure quality inspections nor for checking the work of the contracted maintenance providers. Because of an assignment given by the government after the delivery of the report, the IM has submitted a proposal for changing the way in which infrastructure quality and maintenance work is monitored. In addition, it has presented a tentative program for bringing railway maintenance work back in-house. The latter change was asked for by the government although not recommended by the committee.[6] No further government action on the proposals made by the committee has yet been taken.

[6] The current two-party coalition minority government depends on support from the left party to get its budget through the parliament. The further consideration of bringing maintenance back under direct public-sector control, again performing the activities using in-house resources, was one precondition for the Left Party's budget support.

Much of the committee material is concerned with the precise way in which the IM organises its tendering process and the delimitation of responsibilities between the IM and the entrepreneurs in charge of the respective contract areas. The ministry has not reacted on these comments. This should be seen against a background of the government office being small with limited resources to intervene in the agency's organisation and day-to-day activities. Some 5-10 people are involved in the operational steering of the IM, including both road and railway issues. The ministry has no railway engineers employed and typically does not use consultants for reviewing material from the IM.

The IM is now busy developing a management tool of the type suggested by the committee. It is not clear whether this will suffice to deliver support for decision making except for with respect to a better and more comprehensive database for registration.

4.6 Sweden's performance scheme

The SERA Directive requires member countries to implement a Performance Scheme to incentivize railway operators and the infrastructure provider to improve the performance of the railway system, i.e. to reduce train delays. The IM administers the Swedish version of this scheme.

In July 2014, the government assigned VTI to support the IM in its task to further develop the scheme. The subsequent report (Nilsson 2016) concluded that for three reasons, its present design does not provide the qualities that are necessary to deliver the incentives that are to be met by this type of mechanism. The first deficiency is that information about delays and their causes is incomplete. The most important deficiency occurs when a (primary) train delay have consequences for other trains in the system. Irrespective of the reason for the primary disturbance, subsequent trains may be re-routed and operated over a different path that originally planned. The information collection process of today makes it impossible to register un-planned re-routing of services. While the annual number of disruptions of this nature may be low, passengers and freight customers may at each occasion be severely affected.

Secondly, today's system charges operators and the IM for their own delays but not for the knock-on consequences for other operators because of break-downs etc. of the first operator's trains. This is contrary to the purpose of the system, which is to inform the responsible party about the full implications of a primary disruption by way of making them pay the charge. The ideal way to provide incentives to reduce the risk for disturbances is therefore not implemented.

The third concern emanates from that the IM tenders all maintenance activities. Delays emanating from infrastructure failures and the concomitant performance charges are not paid by the entrepreneurs but by the IM. This stops the appropriate signals reaching the party that is best able to reduce the risk for recurrent failures.

Since the report was submitted to the government, the IM has increased the level of the charges while the structure has not been changed.

4.7 The (absent) multi-annual contract and other means for performance monitoring

The SERA directive mandates Member Countries to establish a multi-annual contract between the government and the Infrastructure Manager. It also opens for the implementation of some other regulatory system that may supplant the multi-annual contract. One of the core purposes of this contract is to highlight the importance of overall cost efficiency. This links directly to the purpose of the present paper; if it can be established that the IM is minimizing costs, it is more reasonable to expect that the benefits of implementing innovations in the way in which investment and maintenance is performed becomes apparent.

Sweden has not implemented a multi-annual contract. The government still believes that the gist of the Directive is addressed. One reason is that the domestic acquis establishes that an overriding responsibility for all public-sector agencies is to ascertain that their operations are implemented in an efficient way. Agencies are also responsible for continuously finding new ways for implementing its assignment (Myndighetsförordningen 2007:515). Another ordinance establishes that the expedience of existing rules and regulations that governs the implementation of the agencies tasks shall continuously be evaluated (Förordning om årsredovisning och budgetunderlag 2000:605).

Moreover, in response to EC Directive 2001/14 on infrastructure costs and accounting – the predecessor to the SERA directive – the government drafted a mimeo discussing the necessity to implement a multi-annual contract (*Faktapromemoria 2007/08:FPM97*). Directive 2001/14 on infrastructure costs and accounting stipulated that Member States are supposed to ensure the gradual improvement of infrastructure and reduced costs for the infrastructure costs. Costs must be reduced without compromising security and the quality of services. At that time, the Directive established that an agreement between the ministry and the IM for a period of at least three years would be one way to implement this responsibility. Except for its focus on cost efficiency, the agreement is supposed to grant the IM a stable source of funding.

One alternative to the multi-annual contract would be to give the regulator the mandate to monitor the IM. However, section 3.3 established that Sweden's regulator is not made responsible for monitoring the IM's performance. The government however argues that since the Swedish IM is a public agency, and not a company, oversight by a regulator is not appropriate. By default, the government is supposed to be responsible for what may otherwise be handled by an independent regulator.

The reference to the government's 2007 position on this issue as well as the interviews that have been part of the drafting of the present mimeo, indicates that several means are used for meeting the requirements. One mechanism is the reference made to the overall public-law framework that public-sector agencies are working under which provides an overall framework for efficiency. Sweden's version of this framework comprises principles for annual funding, requirements for how the money is to be used and the objectives that the IM is supposed to achieve (cf. description above). Moreover, the long-term framework for allocating investment and maintenance resources provides a guarantee for that the IM can be working without concerns over the future availability of resources. The fear of budget cuts will not deter the IM from taking a long-term perspective in the planning of its activities.

In the description of duties of *Trafikanalys,* reference was made to its annual review of the aggregate policy objectives set out by the parliament (section 3.3 and Appendix A). In addition to this review, the government office has established a governance framework for the monitoring of operations and maintenance of state infrastructure, both roads and railways. Based on *Trafikverket's* annual report for 2016, Appendix B provides a summary of this report which includes the following parameters.

Punctuality refers to the railway system's ability to deliver planned travel and transport times, as well as its ability to rapidly provide correct information in the event of disruptions. The target for year 2020 is that 95 percent of all trains shall arrive at the latest within five minutes of the arrival time according to the timetable, and that 80 percent of travelers consider traffic information to be good or acceptable in the event of disruption. Punctuality is demonstrated to be largely unchanged from 2014 to 2015.

Robustness relates to the railway system's ability to withstand disruptions, as well as the IM's ability to handle disruptions, if and when they occur. For this reason, the transport system shall be built and maintained to be reliable when exposed to stress caused by landslides, avalanches, hazardous winter conditions, storms, flooding, accidents, and other unexpected incidents. Increased spending on maintenance and renewal, for instance by replacing existing overhead electricity supply, tracks and switches will increase the system's robustness. Robustness is measured with metrics such as train delay hours owing to defects in the infrastructure. Delay hours have increased on major intercity routes but have been reduced in major cities. In the aggregate, these defects decreased during the year, and robustness therefore increased somewhat.

While these measures have at least some link to the way in which the system is maintained – which is the purpose of the monitoring system – the other parameters are not. This refers to capacity, the ability of the transport system to handle the requested volume of travel and transports. Usability deals with the ability of the transport system to meet the needs of various user groups. Safety is quantified by registration of the number of fatalities and serious injuries in the railway system. Finally, targets include the negative consequences for the environment and health.

To conclude, the system that has been implemented provides poor precision for establishing the quality of the railway system, for the proficiency of the IM to handle it and it is completely silent on the issue of costs for delivering this outcome. What is missing is therefore information about for instance the number of train-stops per year triggered by problems with the infrastructure. This should be further disaggregated for delays cause by switches, tracks, signaling equipment etc., and there is even more disaggregate information available about which components of the respective systems that fail. A complementary statistic is the total time that trains are delayed, including the consequences of a primary disturbance for other trains in the system. Information about how often components must be repaired before that a deviation from target quality levels have affected traffic would provide a systematic update on the frequency of repairs and presumably an input for estimates of the need to replace components that malfunction frequently. All this information resides in the IM's different registries but are never compiled in this format.

4.8 Summary

Sweden's overall way to organise its public sector is based on the use of large agencies and a small government office. One consequence for the railway sector is that the way in which the IM specifies the overall need for investment and maintenance resources, and how it uses the results of these analyses for allocation of resources to different parts of the organisation, is not scrutinised by any outside party. Since the same information is used as a basis for the budget dialogue with the (small) government office, no-one has the know-how in railway engineering and the analytical skills for studying the basis for the evidence provided by the IM. Even if the government does not automatically approve the IMs request for resources, budget cuts are typically not based on explicit analyses but on the government's general need to trade off the demand for resources from different parts of the public sector.

This situation is illustrated by the current universal understanding of Sweden's perceived backlog of resources for track maintenance and asset renewal. Problems in the industry, such as well-publicised train delays, are blamed on an ageing infrastructure and governments have for several years allocated extra resources for maintenance and renewal purposes outside the ordinary budget process. This is a means for demonstrating political decisiveness.

There may, indeed, be a backlog in the replacement of tracks, switches, overhead electricity appliances etc. The only substance that is given for the claim is, however, that the expected life of many asset types has been surpassed. But this is not sufficient for establishing a backlog. The annual maintenance may for years have been appropriate, with components etc. being replaced at the right time. Switches etc. may therefore survive for many more years than expected when they were mounted. The opposite situation could also be true, with insufficient day-to-day maintenance, more trains having used a line and its switches than originally expected, etc. The only way to establish the standard of the equipment is therefore to keep records of the frequency of repairs, of how often switches and other important components of the plant break down and cause traffic disturbances, etc. This type of information is available in the IM's extensive set of databases, but it is not compiled in a format that facilitates a systematic analysis of actual standard. And since no-one is asking for information about how failures affect train serviceability, nor indeed whether a current backlog may cause future costs to increase even more, the necessary analysis shines with its absence.

So, the overall consequence is that the IM establishes its planning procedures without any outside party having the competence to question the conclusions. Even if all demand for budget resources is not met, the government takes the IMs statements of resource needs at face value. This also means that no formalised mechanism is in place for monitoring of cost efficiency.

The current situation is also illustrated by that the spending on track maintenance and renewal has ballooned over the last 20 years or so, much faster than the increase in traffic. The reason for this cost increase is not known.

For an understanding of the way in which Sweden's central government works it should be acknowledged that the IM and the regulator are both government agencies, and one cannot supervise the other. Furthermore, the country's agencies have a long tradition to communicate through the government in matters of a principle, not directly with each other. One consequence is that the propensity of one agency to criticize another is low.

An additional perspective on the un-balanced staffing of ministry and agency is the country's overall tendency of trust. The only reason for that many parts of the public sector can be able to deliver high-qualitative services without excessive costs to the tax-payer is that the information that is provided between ministry and agency is not deceptive. Providing trust-worthy answers to the questions raised by the ministry is, however, not the same as saying that the ministry is asking the correct questions. The arms-length distance and the parties' highly un-balanced skills obviously provides a risk that important underlying aspects that are difficult to discern without proper analysis remain unattended. Many public-sector officials today refer to that a policy decision is taken by elected representatives of the people. But the right of policy makers to take these decisions does not necessarily mean that the decisions are correct, or more precisely; the decision may have been different if the policy makers had been better informed by their experts.

4.9 Competitive procurement – an example of an innovative change

One part of the Big Bang revolution of Britain's railways in the mid-1990s was the transfer of railway maintenance from using in-house staff to competitive tendering. This transmission has subsequently been revoked, meaning that maintenance now is done by Network Rail using its own staff. Today, Sweden, Finland and the Netherlands are three examples of countries that tender all track maintenance.

Banverket was made Sweden's IM at the 1988 separation of infrastructure from traffic. At its establishment, the main part of *Banverket* comprised the infrastructure division of SJ, the previous nationalised, vertically integrated monopoly provider of railway services. In-house resources were used for all track maintenance.

At that time, *Vägverket*, the infrastructure agency in charge of road infrastructure, had a similar organisational structure with respect to the use of in-house resources for maintenance and about 20 percent of all road construction. In 1994, *Vägverket* started its process of gradual transfer of road maintenance from using in-house resources towards competitive procurement. Before the new millennium, all road construction and maintenance had been divested and subject to competitive procurement.

Banverket took a first step towards organisational reform by establishing an internal principal–agent structure. In December 1999, the agency was instructed by the government to consider the possibilities and forecasts for introducing competitive tendering also for railways. The agency submitted its (favourable) report in September the following year; cf. Karlsson & Redtzer (2000). Competitive procurement was subsequently initiated in 2002.[7]

The outsourcing process started by tendering two contract areas that were seen to be simple, which means that they included lines with low traffic intensity and low technical complexity (Espling 2007). Based on the lessons from this cautious start, some two or three new maintenance contracts were put out for tender each year. All contract areas in the country had been tendered as of 2014, when also several contract areas had been re-tendered. Renewals are tendered separately from the 5-7 year contracts for ongoing maintenance and – at least for some of the maintenance contracts – so is also grinding.

Using data for the 1999 to 2011 period, Odolinski & Smith (2016) provide evidence of cost savings around 11 percent; the (few) indicators of infrastructure quality that are available did not signal deteriorating quality. This means that the increase in maintenance costs over this period would have been even higher if maintenance had not been outsourced.

[7] *Vägverket* and *Banverket* was subsequently merged in 2010, now under the name *Trafikverket*.

The way in which the tendering process is organised, and the specification of the responsibilities in the contract between the principal (the IM) and the agent (the entrepreneur submitting the winning bid), is decisive for the outcome of the assignment and for the principal's costs. This section describes some features of one specific contract. Since the details of contracts are continuously updated, it is not clear how representative this particular contract is.[8]

Much of the railway specific equipment that is used is tendered by the IM and made available to the respective holders of contracts. This means that any scale economies in the purchasing of equipment is dealt with outside the contract.

The portal paragraph of the maintenance contract specifies the agent's mission:

"The assignment includes all activities required for maintaining the standard of the railway plant as established in the takeover inspection and to meet other quality standards established by the IM. The expected extent of work is established in the quote for bids and shall be complemented with the entrepreneur's professional assessment of which activities that are required to …

- maintain the status established by the takeover inspection,
- ascertain that the line is operational all days of the year, and
- that the functional requirements established by the IM are satisfied."

The same paragraph also states that the entrepreneur is supposed to ascertain that the lines are available for train traffic per the existing time table for all days of the year.

With one exception that will be further described below, the contract does not specify the entrepreneur's obligations in addition to the above text. Rather than enumerating precisely which activities that are to be performed, the contractor is therefore expected to choose the extent of repair, tamping and other activities that are necessary for "maintaining the standard" of the plant.

However, even diligent maintenance cannot avoid tracks and structure from ageing over time, and more so the more extensive the traffic is. It is therefore – by definition – not feasible to meet the portal paragraph's requirement to maintain "…the standard of the railway plant as established in the takeover inspection". The contract is silent on how this aspect of the assignment is established and monitored.

There is a degree of performance requirement in the formulation, in that the plant shall be of a quality that facilitates the implementation of train traffic as specified in the annual schedule. But even if this target is met, i.e. even if no trains are delayed because of a defunct infrastructure, the ageing of the plant may have consequences for the ability to deliver timely train traffic in the future, for instance because of insufficient preventive maintenance. The contract does not include any explicit acknowledgement of any intertemporal trade-offs in the assignment, i.e. on the consequences for future quality and costs of the implementation (or not) of maintenance by the current contractor.

[8] This description in this section is based on a contract for Mälarbanan covering the 2013 – 2018 period with option for further extension.

The level of maintenance costs is highly dependent on the possibility to get access to, and work on the tracks for a coherent time-period. A standard concern is that this is not possible, i.e. that maintenance must be interrupted so that train time-tables may be operated. While emergency repairs can have repercussions for the time table of trains, all planned maintenance takes place within carefully defined time windows, i.e. time-slots allocated to the entrepreneur. It is, however, not clear from the contract how these windows are established. Consequently, entrepreneurs' bids must be placed with a substandard information in this dimension.

Inspections

Inspections are at the core of the contract between the parties. When the IM sends out a quote for bids, each entrepreneur that intends to submit a bid for the contract must undertake a more or less rigorous inspection in order to estimate the costs for the job. The portal paragraph also establishes that the standard of the railway plant at the date of take-over provides a reference point when service delivery is evaluated. To establish this standard, a takeover inspection is jointly performed by the parties.

The contract moreover mandates the entrepreneur to inspect the plant annually and establishes in detail which aspects that are to be scrutinised. The result of the inspection establishes the status of the plant and which glitches that must be remedied to keep the assets at the required standard. The examination is extensive, in particular for electric and signalling installations.

The entrepreneur is also required to undertake safety reviews to identify activities that must be executed to maintain traffic safety. Based on traffic volume, maximum speed, age, climate etc., each asset class is categorised in one out of five classes, requiring increasingly frequent controls. A switch is, for instance, to be controlled six times per year for the highest class while only once for the lowest priority class.

A distinction is made between the severity of notifications at inspections. An acute (A) note indicates imminent risk for accident or train disturbance and is supposed to be remedied immediately; track closure or speed restriction is to be considered. A V notification is to be repaired within two weeks, etc. The results of the inspections are to be documented in a system designed for this purpose.

There are complementary, mechanical inspections of asset quality. This includes non-destructive testing of rails and rail components, using ultra sound equipment. Quality of track geometry as well as overhead catenary is measured by specially designed trains. These inspections are tendered separately from the maintenance contracts and results are made available for both the principal and the agent.

Payment

The total remuneration for this contract period is SEK 180 million. This is further specified for each year of the period. The following example indicates how the compensation is paid the first two years of the contract.

Year 1: SEK 36 million; no indexation.

- Fixed payment: SEK 28,0 million
- Quantity regulated payment: SEK 6,6 million
- Target cost payment: SEK 1,6 million

Year 2: SEK 35,6 million; the payment is indexed.

- Fixed payment: SEK 27,2 million
- Quantity regulated payment: SEK 6,6 million
- Target cost payment: SEK 1,6 million

The bulk of the payment – in this instance about 75 percent – is fixed. The bidders are aware of the requirements formulated in the contract and makes an estimate of the costs for taking on these responsibilities. The costs for performing inspections and for implementing the repairs that are necessary to meet the set quality targets must be covered by this payment.

However even this component is not completely fixed. When the cost for repairing a failure is “high”, the contractor is paid extra. If the repair costs for instance exceeds a cap of SEK 10 000 – say that it costs SEK 15 000 – the contractor will be paid 5 000 SEK in addition to the fixed payment. This reimbursement rule can vary between contracts, creating different levels of power in the incentives. The same reimbursement rule is used for repairing a defect before it becomes a failure, i.e. for costly preventive maintenance.[9]

Some interventions and repairs are costlier and/or are less straightforward to quantify at the time the contract is signed. A second component of the annual payment is therefore remunerated based on a unit price structure. In the quote for bids, the IM has for instance estimated that 50 concrete, and 200 wooden sleepers can be expected to require replacement during the five-year contract; the entrepreneur’s bid establish that this will cost SEK 4 846 and SEK 1 757, respectively, per sleeper. The payment will be for the actual number of replacements that are required after that the IM has approved each additional change of sleepers.

The third payment component is related to a “target cost”. This refers to costs for handling winter conditions, in particular snow removal. An annual target level is set for these costs. The entrepreneur is paid on a unit cost basis for winter maintenance up to the target level. In this contract, the entrepreneur would receive 40 per cent of any cost savings relative to the target, i.e. the difference between actual and (in the bid) estimated spending. On the other hand, if the target cost is exceeded, the IM only pays 60 percent of the increase to the contractor.
The fixed-price component in the above schedule is getting lower by the year; this is a means for inducing the contractor to improve cost efficiency. This component is, however, also indexed against Net Price Index, meaning that the IM carries the risk for general price changes. Payment for quantity regulated costs is indexed to account for that the price of railway specific materials develops in a different way than for other costs.

[9] Odolinski (2015) analyses the properties of this design.

Penalties

To ascertain that the entrepreneur does not deliver sub-standard quality, the IM has a system for penalties. The following examples illustrate which types of (mal-) performance that are penalised:

Repair reports: Each repair of a failure is to be registered in a database designed for this purpose; the report is supposed to be submitted immediately after the work has been done. A fine of SEK 500 per repair and delay day is levied. It is, however, not obvious how the contract considers defects that may be detected by the entrepreneur's staff and remedied without registration in the database.

Late repairs of registered faults are metered out at SEK 2000 per fault per new day (type A faults, which should be fixed immediately) or new week (type V faults, which should be fixed within two weeks).

The response time is the time between the entrepreneur becoming aware of a problem that requires immediate action and the implementation of the repair. There is a fine of SEK 5 000 for each additional hour of response time violation.

Train delays. For the delivery of a well-functioning infrastructure, timeliness of train services is obviously of utmost importance. In various ways, all planned maintenance takes place within carefully defined time windows, i.e. time-slots allocated to the entrepreneur. If these windows are violated so that trains are delayed, the fine is SEK 15 000 for each new ten-minute-period for each train that is affected. The total fine is capped at 10 percent of the annual value of the contract.[10]

There are some systematic comments to pass on the presence of barriers to innovations within the IM's railway activities. First and foremost, railways in all countries are built and maintained based on a wealth of administrative rules and regulations. The acquis concerns both how activities are to be implemented as well as which equipment (rail weight, signalling system, etc.) that is to be used.

While this will inevitably delay the innovativeness of the railway industry, there are several motives for the use of administrative rules. One is related to safety concerns and the necessity to ascertain that neither the equipment nor the way in which maintenance is done cause accidents. Because of extensive certification procedures that are supposed to handle these concerns, new components cannot be start being used on short notice. This makes the whole industry less innovative than many other sectors of the economy.

The IM's central office, and its specialised staff (signalling, electricity, tracks-fastenings-sleepers etc.), has extensive control over the administrative framework. This includes research using in-door experts as well as the funding of work at universities and research institutes. Another strategic means for the IM relates to on-going contacts with manufacturers, inter alia to ask for new qualities of the equipment. It goes beyond the scope of the present inquiry to assess how well this market interaction is working.

A second challenge for innovativeness in maintenance relates to the design of maintenance contracts and its consequences for testing new ideas. The previous section has described several features of

[10] In personal communications, it is indicated that the IM rarely meters out this penalty. The reason is that repairs may surpass the ex-ante time window because of late trains that makes it impossible for the entrepreneur to start working at the promised time. It is therefore seen to be inappropriate to fine the agent for this knock-on effect.

these contracts. There is obviously some scope for entrepreneurs to develop use their resources – primarily their staff – in new ways for handling the maintenance, again within scope of the administrative restrictions. Any cost savings that will not jeopardise quality will benefit both the principal and the agent to the contract. The main concern in this seems to be in the ability of the contracts to handle intertemporal aspects. While contracts are signed for five to seven years most technical components of the infrastructure as a whole last much longer. This means that the fragmentation created by separating infrastructure from train operations is not solved by a transfer to competitive procurement, less this is explicitly dealt with in the contracts.

Except for maintenance, the IM annually spends large amounts of money on the renewal of tracks, signalling and the power supply system. For this purpose, it is necessary to use Life Cycle Cost information in order to prioritise track sections, switches etc. that are at the end of their economic life. This information is today not available. The prioritisation of renewal interventions is, however, a challenge which must be handled outside the contracting process per se, since it is the IM that must make the priorities. Again, the fragmentation from vertical separation remains.

The benefits of coherent time frames for renewal activities, inter alia by closing a line for a period of (summer) time, is another aspect that affects the cost for these tenders. As of today, this seems to be handled by the IM separately from the quote of bids. A discussion is presently under way within a VTI research project that considers the possibility to make the timing of maintenance and investment activities an explicit component of the bid for road (re-)investment projects. This would then be handled by way of delay fees, an extended version of the concept of lane rentals in the road sector. In principle, it is feasible to extend this approach also to railways.

4.10 Examples of challenges to dynamic efficiency in a vertically separated railway industry

Commercial firms face a constant pressure to develop their way of working, both to stay in business and to win new customers. The analysis of how competition force suppliers to be prudent with costs in the short run, as well as to develop innovative ways for remaining in the market in the long run is a core topic for economics text-books.

But the way in which competition fosters static and dynamic cost efficiency differs between industries. One obvious feature of railways to make them differ from many other productive activities is the significance of scale economies. A conjecture is that this quality provides a background motive for the separation of infrastructure from train service operations. Compared to the traditional railway organisation, this fragmentation creates an additional idiosyncrasy of the market. This makes it necessary for the IM to acquire information about the impact on train operations of different approaches to maintenance and renewal of the infrastructure.

The Swedish context adds a further dimension to this challenge. Since the IM does not use in-house resources but tender's maintenance in competition, the contract with an entrepreneur must be designed to pick up the consequences of different types of interventions on traffic and on the ultimate beneficiaries, the travellers and freight customers. But the contract must in addition be designed in the awareness of that its clauses – the allocation of responsibilities and risk, the triggers of maintenance, the length of the contract, etc. – have consequences for the way in which profit maximising entrepreneurs act. Contracts must be crafted with this in focus.

Section 5.1 considers one dimension of this challenge which relates to a problem that has been observed also for contracts in other parts of the economy; it is referred to as a common value risk. Track user charges pay only for a small share of maintenance costs, meaning that resources for track maintenance and investment is highly dependent on complementary resources from the general budget. Section 5.2 addresses "the Valley of Death", another concept that has been coined in completely different applications. It relates to a specific challenge for dynamic efficiency – for being innovative – in a railway sector which is heavily dependent on public sector funding.

Section 5.3 considers one specific example of the challenges provided by maintenance contracts covering a shorter period than the asset's expected life-length. The example concerns the use of information about track quality for undertaking maintenance activities that in the first place has benefits outside the termination of the contract. Section 5.4 discusses the incentives for developing better information about the ground condition at a construction site. The focus is on the incentives for innovation when several parties are involved; the IM representing the interest of the public, the companies that survey the ground as a basis for preparing drawings and construction plans and the entrepreneur subsequently implementing the project.

While the previous section focused the responsibilities of the IM for designing contracts with the incentives of entrepreneurs in mind, the subsequent examples have the opposite subject in focus: How may entrepreneurs react in the situations that are described, given the contractual framework at hand?

Common-value challenges in maintenance contracts

The costs for railway maintenance is largely a consequence of the track standard; the more traffic and the older and the more worn-down the assets are, the costlier is the maintenance. When a contract is to be (re-)tendered, a crucial component of the IM's quote for bids is therefore to provide a comprehensive description of the quality of the plant; in combination with manual inspections, this information is vital for the preparation of the entrepreneurs' bids.

Bidders are therefore fed with information about mechanical (track measurement trains etc.) registration of standards. In addition, the incumbent entrepreneur is mandated to provide a detailed registry of all maintenance activities that have been implemented. Registration of frequent repairs does per se indicate poor quality, at least if the specific asset that is repaired (a switch etc.) has not recently been replaced.

Measurement of both track quality and the quality of the plant as a whole, is an inherently incomplete science. But with access to the same information about quality indicators, both the IM and bidders are supposed to face the same uncertainty about the underlying quality of the assets. There is, however, a risk that the situation is more complex than so. This relates to that the incumbent entrepreneur may know more about track quality than the competitors in the race for a new contract. One reason is that many dimensions of quality that are difficult to measure still are visible for those that regularly is working on the assets. In addition, the incumbent may have made repairs that are not registered in the systems that are supposed to be used for this purpose.

If comprehensive and correct information is not common to all bidders, the playing field is distorted to the benefit of the incumbent. If outside bidders regularly underestimate the amount of work that is necessary to meet the targets set in the contract, their bids are lower than the incumbents bid. Submitting an overly optimistic bid increases the chances to be awarded the contract. The obvious risk is that this results in a negative financial outcome of the contract.

There are several examples of conflicts between the IM and an entrepreneur where the latter has gone to court to enforce higher payment. The basis of court cases is that the entrepreneurs claim that the track standard was never properly accounted for in the quote for bids. To keep the track in a standard that permits uninterrupted traffic, the winner – the entrant – has therefore had to spend more on maintenance than assumed when submitting the (winning) bid. In one example, the process was settled outside court by the IM agreeing to pay the entrepreneur an extra SEK 145 million.

This is an example of a winners-curse type of problem which is seen also in many other sectors of the economy, most notably in the bidding for oil drilling licenses. The textbook recommendation is to use open bidding for reducing the risk for winners curse rather than the one shot, sealed bid format used in all standard procurement contests. This means that entrepreneurs compete by submitting bids in a common, open domain. Having observed one bid level, it is feasible for a competitor to enter a new, lower bid. This goes on until no-one is interested in changing their bids any more. The benefit of open bidding is that participants in the contest observes when competitors drop out. This may make the stayers aware of that their cost assessment is more, and possibly overly optimistic. The open process therefore provides bidders an opportunity to update the cost belief and will then reduce the risk for making losses.

The Valley of Death

Many entrepreneurs have problems to finance the development of innovative ideas into commercial products since the financial markets finds the probability of success poor. The same challenge is faced by enterprises that – after a start-up period – need new capital for expansion and for supplying goods or services to a larger market. The gap between interesting ideas and commercial implementation is sometimes referred to as the Valley of Death.

The reluctance of lenders to approve these loan applications goes back to the absence of collaterals and the substantial risk for default. The value of an innovative idea may, however, be substantially higher in the development of goods and services with public good qualities than in activities developing more standard private goods for consumers or producers. This is so if an entrepreneur develops a gadget that may benefit a larger audience than the direct users. Developing innovative components to be used in tracks or in any part of the infrastructure plant will thus benefit many trains and their ultimate customers, passengers as well as freight clients. Such gadgets may pass the BC test if considering their wider use, while not in a narrower perspective of train operator benefits.

But a further dilemma for development in the railway industry is that there is only one potential buyer of a new gadget in each country, i.e. the IM. Commercial lenders would therefore have to rely on the ability of a state-owned monopolist to recognise the benefits of the innovation. Moreover, the lenders must believe that the IM is rational in so far as an innovation that meets the BC test also has a reasonable chance to be implemented. The sum of these caveats implies that new ideas that would benefit the maintenance of infrastructure have a lower chance to receive funding than B2B or B2C projects.

One example of a project of this nature is provided by an entrepreneur that recently has won innovation prices for developing the blue-print for a new type of switch. Winters with much snow and with frequent swings between temperatures above and below the freezing level increase the risk for that the moveable parts of a switch get stuck. This would make the signals shift to red with severe consequences in the form of train stops. If the new design reduces the number of emergency interventions by the maintenance contractor, the entrepreneur would save on costs. When the contract is next to be re-tendered, it would also drive down the bid price. But in addition, train operators and their customers would benefit from a lower number of train stops.

This entrepreneur must raise some SEK 30 million for building a prototype switch that can be tested in operation. This process has now stalled. While there is no guarantee that this example of a new type of switch saves more that it costs to develop and to install, that the hurdles for trial-and-error in this line of business are higher than in many other industries. It is also obvious that – given the current design of maintenance contracts – individual entrepreneurs have limited incentives for financing a gadget. Even if it could be demonstrated that one new switch would reduce costs for winter maintenance by so much that it would meet the BC test over its life cycle, there is no guarantee that the holder of the contract would retain it also for future periods.

Track quality measurement

Implicit in the description in section 4 of the design of track maintenance contracts is the necessity to detect infrastructure deficiencies. The earlier that quality is observed to be deteriorating, the faster can remedial (preventive) action be taken and the lower is the maintenance cost. Vice versa, emergency repair of track failures is costly for the entrepreneur as well as for train operators and their customers.

Today, two approaches are used for detecting deficiencies. The manual method means that the entrepreneur's staff – following the contract requirements – undertake maintenance or safety inspections. In addition, when making repairs or when travelling along the tracks, the staff may detect problems outside the mandated inspections. The other approach relies on information from a measurement train that records information about the position of rails. Measurement trains record the quality of main lines a couple of times per year while tracks with less traffic are measured less often.

NeTIRail, the project that has triggered the discussion of incentives, inter alia studies the benefits of obtaining more regular and precise information about track quality. Two strategies are considered. WP 4, task 4.2, studies the installation of on-train monitoring equipment referred to as ABA. Fitting this equipment into existing rolling stock is relatively costly but would provide detailed information about the train's vibration. Recording this information from a train set that uses the same tracks several times each day or week, and using GPS-type of positioning of the vehicle, would make it possible to detect if the vibrations are increasing at certain spots along the line. Task 4.3 addresses the possibility to use an alternative or complementary device, namely the installation of a smart-phone for registering the train's vibrations. This would be cheaper than the installation of a permanent recording device but would, on the other hand, provide less precise information.[11]

Either of these two innovative devices would make it possible to register train vibrations during each leg of a train's movement. The link between readings and geography would be feasible to identify if, when and where vibrations gradually increase. This may in turn make it possible to detect track irregularities at an early stage of a deterioration process. The prime benefit of this would be to facilitate repairs at an early stage and to reduce the risk for temporary speed restrictions that may be necessary to eliminate the risk for severe problems.[12]

From the dynamic efficiency perspective, it is necessary to consider the entrepreneur's incentives to

[11] At this stage, it is not yet clear whether the innovations considered by NeTIRail pass the Benefit-Cost test.

[12] A similar idea was suggested to Sweden's IM almost 10 years ago. The background is that all modern rolling stock (railcars) have equipment that records train vibrations as a standard. The purpose of the kit is to provide input for maintenance of the rolling stock and to make repairs at an early stage to reduce the risk for costly breakdowns. The IM was informed about the possibility to use this information as an input for detecting also infrastructure defects in about the same way as described above; cf. Ekman et al (2006) and Holst et al (2012). For doing so, the IM would have to negotiate access to the information with the owners of the rolling stock and possibly also to pay for the transfer of data. This has not led to any action by the IM.

start the development of this type of equipment. The initial outlay for doing so would be the costs for any type of equipment that could deliver the type of information that is relevant for track position registration. This would also include a process of validating measurement data to ascertain that it does provide relevant information for early detection of vibrations. For full scale implementation, there may in addition be costs related to retrieving and processing the information to make it relevant for triggering decisions about interventions. In the second bowl of the scale lies the savings in maintenance costs facilitated by the improved knowledge about track quality.

It is obvious that the probability of this improvement to be implemented by the entrepreneur faces the same issue of fragmentation as discussed before. Even if collection of relevant information of this nature would enhance efficiency in the long run, the risk is obvious that maintenance contracts that cover at most seven years would be insufficient for motivating any maintenance entrepreneur for funding the innovation.

The design of the contract could affect this conclusion, for instance using carrots linked to a gradual reduction of the number of train delays related to track-quality deficiencies. The current design of the contract as described in section 4 is not sufficient for contributing to this.

It is obvious that an IM instructed to maximise social welfare in the use of its resources, should ascertain that the type of equipment that now has been described should be developed in the case of a positive outcome of the BC test. This would, however, throw the responsibility back to the IM in its role of guardian of the public interest. Again, the design of tendered contracts does a poor job in ascertaining the development of innovations.

Geophysical measurements

The final example of hurdles to innovations in the railway sector is taken from infrastructure investment. Major investment projects are preceded by pre-studies, including analyses of the underground conditions where the new tracks are to be built. Constructions on sites where ground conditions are favorable is much cheaper than if a project must be built over wetlands or if a tunnel must be blasted or drilled. The ground conditions may even be important for where a project is located, i.e. difficult ground conditions may motivate detours that require more tracks but that makes the construction per meter cheaper.

Since the ground conditions strongly affect the costs for building new roads and railways, and since infrastructure is more geographically extended compared to the construction of buildings, extensive assessment of the geology is necessary. High-quality ground surveys are also important since mistakes that are detected when a project is under way is typically costlier to handle than if they were detected at an early stage of the process. Except for that the adjusted design is costlier than the original, the IM and the entrepreneur that has won the original contract must negotiate change orders. The entrepreneur's negotiating position is then improved compared to when different entrepreneurs were involved in the original bidding contest. This may further increase the cost to the IM of the adjusted design of the project.[13]

[13] Anecdotes about major changes being the major source for profits in infrastructure construction abound, i.e. that the winning bid cover costs but not much more while changes and additions generate surplus. Xx (2014) provide some evidence for this for Californian pavement contracts.

One of the most important reason for cost overdraws in construction projects is related to that plans must be adapted to geotechnical problems that are detected when a project is under way. The ground survey has then failed to identify the actual quality of the ground. A misjudgment of the rock type and quality has a strong impact on the choice of method for implementing a project; it may even question its relevance, as seen for example in Sweden's Hallandsås project.

Infrastructure projects have for a long time been dominated by unit price (Design-Bid-Build) types of contracts; this has been documented in previous sections. This contract design means that the principal must carry the costs for cost overdraws because of unforeseen events during the construction phase of a project. One important reason for using this allocation of risk between principal and agent is that the opposite contract type, the fixed price (Design-Build) contract, puts all risk on the agent. Therefore, tenders for fixed-price contracts force the entrepreneurs to add a risk premium to the bid that they submit. This increases the average cost for implementing projects when this contract form is used.

But irrespective of which contract design that is used, improvements of the quality of ground surveys at an early stage of the planning and tendering process project would obviously be important for both construction costs at large as well as for reducing the uncertainty related to the actual conditions at a construction site. Ground radar has become commonplace because of the possibility of fast data acquisition and processing. Its penetration is however limited in environments such as shallow water table or clayey areas.

The last few years have seen a development of complementary geophysical methods that can be used for measuring the quality of rocks. Significant improvements in the time for acquiring and processing electrical, seismic and magnetic data are now making these methods able to contribute to improved planning of construction projects in mountainous terrain. Further validation of geophysical assessment methods would reduce construction risks for projects built under these circumstances.[14]

It is obviously difficult to conceive of the development of this type of equipment within a commercial perspective. Consider for instance entrepreneurs bidding for a DB project where the costs for geotechnical surveying is part of the costs that provides the basis for the bid that will be submitted. Assume, for the time being, that one bidder has taken the risk to develop and use new geophysical methods for surveying the ground. Except for having spent extra on this development and information acquisition, the only thing that would affect the position of this firm would be that the ground provides more challenges than the competitors realize. But this would mean that the innovative firm would submit a higher bid than the competitors, reducing the probability of winning the contest for the contract. The novel approach would not be used under this framework.

The same challenge would prevail under a DBB framework. It would then be the consultants that bid for the pre-study, including the geophysical survey that would have to consider the use of the new equipment. But again, using the novel approach would most probably be costlier than if standard surveying is applied; the innovator would – ceteris paribus – have to place a higher bid than its competitors, losing the contest. The new method would not come to be used.

[14] A complementary benefit of these methods would be to monitor the use of the existing infrastructure, for instance to protect aquifers and more generally for the monitoring of vibrations.

The bottom line is again that the IM has the ultimate responsibility for ascertaining that new methods, here in the form of geophysical instruments for improving information about rock quality, are being developed. A first step in this is to consider the probability for that investing in the development of the method would reduce the construction costs of projects in mountainous terrain. There is prima facie reason to believe that the equipment would pass this test; even if the difference between information provided by traditional and the new approach to surveying may be small in many projects, the benefits could be very high in situations where the extra information identifies severe challenges.

4.11 Concluding comments

The trigger for the present inquiry is the work within the NeTIRail-INFRA project. The centre-piece of this line of work is that of several engineers and railway specialists developing some ten proposals for new ways to measure track quality and for changing the way in which railway infrastructure quality at large – also including the catenary – is maintained. An economic appraisal of these improvements is supposed to establish whether there is reason to go further towards implementation; the question is if the innovations pass the Benefit-Cost test.

But an affirmative nod to implementation would not be enough. This section has therefore considered incentives for innovations to be developed and implemented in a system where infrastructure and train operations are separated: Would an Infrastructure Manager have reason to be innovative in a system where at least some of the benefits of new approaches to investment and maintenance accrue to one or more other organisations, i.e. to the train operators?

The common understanding is that one draw-back of the switch from a traditionally integrated industry to vertical separation is this fragmentation. This can, in principle, be overcome by instructing the IM to account for social welfare, i.e. to consider all costs and benefits of investment and maintenance activities, irrespective of where they appear. One obvious problem with doing so is that relevant information may not be available about effects for outside organisations. Another challenge is the IM's incentives to implement the welfare maximising policy, if it could be identified.

Neither of these issues have been addressed in this paper. Focus has rather been on the above challenges in Sweden which differs from many (while not all) of its European peers in at least one aspect: All construction and maintenance is tendered from commercial entrepreneurs.

The main conclusion is that this difference has only limited impact on the generic issue of fragmentation and its consequences for dynamic efficiency: It is still the IM that has the ultimate responsibility for ascertaining an innovative railway industry.

"The market", i.e. the entrepreneurs that build and maintain the infrastructure, however adds to efficiency in one important way. This is so since the competition for contracts force the firms to develop their proficiency in doing what they are supposed to do. This does, however, to a very small extent carry over to do things in innovative ways. Except for that railways – for very rational reasons – are circumscribed by rules and regulations that control for safety and compatibility of equipment, another reason is that the standard contract formats reduce the degrees of freedom in implementation of the activities. Moreover, contracts for both investment and maintenance activities are typically for far shorter periods than the economic life of the installations. This is another reason for that the IM still is responsible for the development of any infrastructure assets.

The review has also concluded that the culture of cost control is not strong in Sweden. The system depends on a government agency (Trafikverket) appraising options using cost benefit analysis and pursuing efficiency in line with its declared objectives. But no formal regulatory system or multi annual contract sets targets for improved efficiency.

There is no outside competence for monitoring the IM's activities and to ascertain cost control. One consequence is a steady increase in maintenance costs and claims for a maintenance backlog which lacks substance. In addition, cost increases don't manifest themselves by way of red numbers in a financial statement but by postponing planned activities. In the economics literature, this is referred to as a soft budget constraint.

In an industry where users are supposed to pay for the full costs of at least maintenance of the infrastructure, train operators would have a strong incentive to monitor and criticise the IM if charges would be seen to be excessive. Sweden's track user charges are however very low, accounting for some 10 percent of maintenance costs. This further reduces the interest from outside parties to take an interest in the costs for providing infrastructure services.

Comparing Sweden with other European countries, in Sweden infrastructure decisions are taken by an agency which is given the maximisation of net social benefit as its objective, and likewise rolling stock decisions for franchises are taken by an organisation that is not time limited, like the franchisee, and therefor may be expected to look at lifetime costs and benefits in its decisions. The ability to pursue these objectives is not constrained by short run regulatory targets. On the other hand, the relative lack of pressure on costs may remove what may be an effective source of pressure to innovate.

On the other hand, contracting out of maintenance work may provide strong pressure on costs but make innovation more complex by splitting the function between the infrastructure manager and contractors and putting in place contracts which are relatively short term and with limited flexibility to change the way things are done.

5. Germany

5.1 Introduction

In Germany, the main passenger and freight operator remain part of the same holding company as the infrastructure manager (DB Netz). They are regulated by separate regulators for safety and technical issues (*Eisenbahn-Bundesamt*, *EBA*) and economics (*Bundesnetzagentur)*. Long distance passenger and freight services are open to open access competition. Regional services are franchised by the federal states, sometimes in the past by direct award but in many cases, and for all new contracts, using competitive tendering.

Incentives are provided to DB Netz through the following:

5.2 Multi-annual contract

This provides for government co-funding of renewals expenditure (currently €4.5b p.a., including DB Netz´s own funds) in return for which DB Netz commits to achieving a quite sophisticated set of quality standards. It is the safety regulator who is responsible for checking that these targets have been met. Also, DB Netz commits to spending at least a certain amount of its own resources on renewals (currently €0.1b p.a.) and maintenance (currently at least €1.6b p.a.). There are financial penalties on DB Netz AG, DB Station &Service AG, and DB Energie GmbH for failing to meet these quality standards (also incentives to managers via bonuses). DB Netz argues that the contract is quite demanding as it requires a steady improvement in quality without a corresponding rise in government funding. If more spending on renewals or maintenance is needed to fulfil the quality standards, this must be covered by DB's own funds. This mechanism, together with the above-mentioned penalties, yields clear and strong incentives for infrastructure management.

Under the first multi annual contract, covering 2009 – 2014, quality targets have been met at all times. Under the second one, covering 2015-2019, quality targets have been intensified and extended. Currently, DB Netz has been penalised for the last two years for failing to meet individual target.

The federal government as well as DB Netz believe that even higher quality should be aimed at in future contracts, and funding-requirements are likely rise to solve the remaining backlog of renewals. The multi-annual contract as such is regarded as powerful by the entire industry and even seen as a role model for other modes of transport.

5.3 Incentive regulation

A new law concerning rail regulation was introduced in 2016, partly to make it consistent with EU requirements. Under this the Network regulator, Bundesnetzagentur, has extended powers and competences, including the newly gained responsibility from the safety regulator for monitoring compliance with unbundling requirements, as well as ex ante regulation of access charges by newly formed decision chambers.

In principal, all maintenance and operating costs of DB Netz AG should be covered by track access charges. The regulator calculates allowable expenditure as current expenditure plus a price index minus a productivity index (the average for the German economy as calculated by the government's economic advisory council). Charges must not produce revenue in excess of this at traffic levels of a defined base period (price cap regulation). Thus there are incentives both to cut costs further and to increase revenue by attracting more traffic in that the IM may keep the resulting benefits until the next periodic review (5 yearly).

This process will start in 2019. In 2018 a new set of infrastructure charges based on direct cost plus mark ups will be introduced. By law, charges must be based on train km and charges for regional services must be equal to the growth rate of public grants for regional services, i.e. 1.8% per annum. Mark-ups are determined by Ramsey pricing. But given the lack of information about the contents of trains and about elasticities, the segmentation is crude. Mark ups are highest for peak ICEs.

As noted above, charges are based on train km, with differentiation by type of train, type of route and quality of the path. Charges are not differentiated by gross weight or other aspects of track friendliness, so they give no incentive to train operators to use track friendly rolling stock. Germany does not currently have a performance regime, the Regulator having regarded the scheme previously brought forward by DB Netz as not fit for purpose.

During the legislative process, it was commented that the multi annual contract and the regulatory system are not totally consistent in that the regulatory system incentivises reductions in maintenance costs but the multi annual contract requires at least a certain minimum level and might require increasing maintenance in the future. To solve this inconsistency, the new regulation law allows for deviations from the above-mentioned price-cap induced by the multi-annual contract, if approved by the regulator.

Irrespective of the incentive regulation, the sector is currently appealing for the government to deviate from full cost coverage by providing funds to halve track access charges for freight traffic in order to improve the competitive position of rail.

5.4 Investment planning

DB Netz, alongside other sector representatives, also contributes input to the development of the government's long term multi modal investment programme. The approach emphasises dealing with bottlenecks as a result of forecasts of traffic growth but includes new stretches of high speed rail where these are an efficient way of dealing with these bottlenecks, taking account of time savings. Schemes accepted for this receive a contribution from government large enough to yield a positive business case for DB Netz. The DB Netz plan is called Network 2030, and is based on a Taktfahrplan (interval timetable) approach to the timetable. This has strong political support.

5.5 Conclusions

Clearly there is a risk that elements 1 and 2 place emphasis on short run cost cutting rather than long run life cycle cost minimisation. A number of factors work against this:

- the performance regime provides long run incentives to improve performance

- so do the mark-ups on direct cost of track access by making it profitable to attract more traffic.

- The holding company monitors activities of its subsidiaries to make sure they are consistent with the long run interests of the industry and DB AG as a whole. In this context, however, unbundling regulations require that DB Netz remains fully independent as far as decisions regarding track access charging and capacity allocation are concerned. The alignment of incentives in terms of financial sustainability, proper investment and quality of the infrastructure should lead to better management in favour of the entire industry, compared to a setting where DB Netz would decide investment on a pure stand-alone basis.

An example of this sort of benefit might be the operation of 750m freight trains, which requires investment by DB Netz in longer passing loops but will reduce costs for train operators. Another is the upgrading of the track for higher speeds, removing the need for train operators to incur the addition al costs of tilting trains.

Even more, ERTMS requires a systemic view, since investment by train operators as well as infrastructure managers are required to reduce costs and raise quality on both sides. Due to a complex allocation of costs and benefits and a high necessity of co-ordination, individual decision making would more likely hold-up efficient investment.

The holding company also is responsible for long run strategy and research. It prepares demand forecasts and scenarios, and leads on innovation (e.g. automation and digitalisation), including providing funding for start-ups in this area. There are also some specific government funds to encourage innovation in particular areas (e.g. noise reduction). These funds are open to all train operators. Rolling stock investment is mostly by the operator or by the authority tendering the services. Both have an incentive to consider life cycle costs.

Essentially because DB AG is very much concerned about its long term success, interviewees did not perceive a particular problem of emphasis on short term gains at the expense of long term. This seems to be partly an advantage of the holding company model preventing DB subsidiaries from pursuing their own interests at the expense of the group as a whole.

5.6 Acknowledgements

We are grateful to Peter Abegg, Markus Ksoll and Wolfgang Bohrer, all of DBAG, for information, and to participants at the Workshop on Financing in the Railway Sector at the University of Giessen in May 2017 for further discussion of the issues facing the Railway Sector. Responsibility for the final text is however solely our own.

6. France

6.1 Introduction

European Directives and Regulations deeply changed the French Railway landscape. European vision changed the level playing field in rail.

Historically, the TGV system was a great driver for innovation for railway industry in France. Its placing into revenue service, in 1981, affected not only rolling stock or infrastructure design; it also re-shaped railway product marketing and revolutionized the fare system with the subsequent introduction of compulsory reservation and yield management (project SOCRATE, delivered in 1993).

The innovation period went through different phases, from system design (sixties and seventies) through development and return from experience. The latest, most significant breakthroughs were probably the double decker design and the advanced yield management system (both appearing in the nineties), while the overall transport system concept remained stable over the years.

Other innovations were shelved (e.g. advanced container and mobile load management, replacing the concept of wagon shunting yards by load shunting yards) or occupied market niches (e.g. roadrailers). Further innovations such as ASTREE, an advanced, satellite-based signalling system, the basic design of which took place in the late eighties, became inspiration for ERTMS.

It could be argued that disincentives should be the real subject for discussion, rather than incentives: that is, not all innovative ideas translate into reality. The post-mortem analysis of failed projects might provide useful lessons, if confidentiality concerns, or mere pride, were not an obstacle.

6.2 Stakeholders

SNCF is currently divided in 3 'EPICs' (Établissement Public Industriel et Commercial): the EPIC SNCF, the EPIC SNCF Mobilités operating passenger and freight trains, and the EPIC SNCF Réseau, the French Infrastructure manager. Freight market is open for all operators since 2006, and passenger high speed lines will be opened from 2020.

From 1997 to 2015, the infrastructure manager was RFF; RFF role was however limited to strategic and financial planning, with all operation and maintenance of the network being delegated to SNCF Réseau.

With respect to the French railway institutional organization, ARAFER (since 2016; previously ARAF) is the French regulator for, inter alia, rail activities. This independent public institution was created by the end of 2009. ARAFER oversees the access conditions and the levels of access charges to the infrastructure. This evolution is in line with European law, and the French legislator further empowered ARAF during the French rail reform in summer 2015. To ensure progressive network opening to competitors without discrimination, which is its fundamental mission, ARAFER has the authority to enforce conformity to competition rules, including technical and administrative conditions to access and operate the network. The French National (rail) Safety Authority EPSF (Établissement Public de Sécurité Ferroviaire), created in 2006, gives the green light to operate lines as soon as the conformity with European and national regulations is proven; EPSF checks the coherence and safety of the system, and contributes to the interoperability of European networks, thus establishing a fair and level playing field for all stakeholders.

DRN (document de référence national) is the French legal text transposing European Directives and referencing TSIs. AMEC (autorisation de mise en exploitation commerciale) is the authorization to use subsystems for commercial purposes.

New organisations managing and funding research & innovation such as IRT Railenium, and competitivity clusters (Moveo, Nov@log and I-Trans) were added to the existing PREDIT and IFSTTAR. Transposition of European Directives[15] in French law (April 2016) led to the innovation partnership (partenariat d'innovation) signed between SNCF and Alstom in August 2016 for the 'TGV of the future'.

6.3 Incentive mechanisms

Incentives for innovations are provided by:

Direct subsidies: The State is contributing to SNCF Réseau's budget for infrastructure upgrade and investment, as well as combined transport subsidies.

Complementary subsidies: projects approved by research foundations such as Railenium (see below) become eligible to co-funding by the French state agency for research (ANR, Agence Nationale pour la Recherche), where the state / private share of costs is 50/50. The approval implies that the research purpose lies within the objectives of the foundation.

State guarantee: by virtue of its status, the SNCF EPIC has an implicit State guarantee; in theory, a creditor of an unpaid EPIC could turn to the State. In the French case, this status contributed to attractive financing conditions, which had helped the company to invest in innovative infrastructure. State guarantees may however no longer be compatible with European law, where distortion of competition could result (e.g. in the case of railway undertakings).

Track access charges amounted to around 5.5 billion of Euros per year and need to compensate direct running costs. The pressure to reduce the access charges and stimulate more demand in order to reach the target revenues leads SNCF Réseau to work for reducing its costs and therefore introduce innovative methods (techniques and technologies). This pressure is expected to increase when "competition on the tracks" will be extended to passenger services, provided there is spare capacity to sell, i.e. mostly on the conventional railway network. However, track access charges differentiate by type of route and peak versus off peak, but there is little differentiation by vehicle type and therefore no incentive to employ lighter more track friendly trains.

A possible evolution would therefore be the linking of track access charges to axle loads, an orientation favoured by ARAFER. The link is sensible, in the sense that static axle loads determine direct costs to a significant extent (all other things being equal). Side effects put aside, such an orientation could work as an incentive to further reduce axle loads (a long-lasting trend in Japan for

[15] Directive 2014/24/UE ; Directive 2014/25/UE ; Directive 2014/23/UE

instance); in the worst case, it would be a simple tax on double decker runs, as effects would be immediate and solutions not at hand.

A contractual framework between Regional Transport Authorities and the operator sets a number of expected performances, among which punctuality of trains. Such requirements also put infrastructure design and management under pressure, as most regional lines are mixed-traffic ones. To this end, more innovations approaches are required within the organisation.

Open data: SNCF probably was the first company to launch an "open data" initiative in France, back in 2012. SNCF consolidated its leading position in this field by opening up 200 datasets, published on the platform data.sncf.com. Open data fosters involvement of innovative start-ups and tech companies, generally improving user experience.

R&D tax credits (crédit d'impôt recherche): The French railway sector have benefited from this research-related tax cut to further invest in innovation programs and projects.

Innovation labs, awards and partnerships: SNCF consolidates innovation in technology and new business models, by creating partnerships with industry players from different sectors (including Hyperloop One), and academics among which there are major research laboratories of international dimension. It also announced recently an "innovation partnership" with Alstom, following a competitive tender. The goal is to design a new TGV model.

Public-Private-Partnerships Procurement: One of PPPs' expected main benefits is enabling private sector innovation for public service delivery. In particular, availability based PPP contracts are well adapted to increase innovation by setting the relevant expected performance rather than specifying the technical solutions. Additionally, PPPs are seen, by investors, as an opportunity to access relatively cheap financial resources for traditional procurement (example: LGV SEA, the southeast high-speed line).

6.4 Possible shortcomings or unexplored areas

Public satisfaction monitoring and civil society consultations mechanisms, allowing to collect requests in social media, constitute an incentive for innovating and responding to the demand. In that respect, a consistent, continued set of quality indicators (quality in the sense of "response to implicit or explicit needs of the customers") might however be missing, both at French and at European scale.

"Common quality indicators" as a pendant to "Common safety indicators" are however more of a challenge, as quality is inherently subjective. Also, quality criteria may evolve with time and in consideration of what the competition has to offer; the same applies to safety.

Transfer of responsibilities from railways to suppliers

A progressive transfer of system knowledge and design responsibilities from SNCF to supply industry has taken place since the mid-nineties, first affecting product design (rolling stock, then signalling and interlocking), and progressively extending to system design.

The rationale for this shift, requested by the supply industry and supported by ministries, was the belief that suppliers would become more successful on the international market by proposing products less tailored to the domestic market due the strong intervention of the dominant domestic consumer, SNCF.

The inherent risk was that the direct feedback from operations into design, internally organized and enforced by SNCF along the whole lifecycle of the products, could not be fully replaced by the limited information exchange during guarantee periods or via aftersales services. In France, maintenance services of electromechanical subsystems are mostly in the hands of the operator, with limited subcontracting.

The ex-post analysis of this fundamental shift, that also took place elsewhere in the world, has not been conducted to date. It would require efforts largely exceeding the perimeter of the present study.

Performance regime

The absence of performance regime between RFF and the operators (SNCF mobilités, or private freight transporters) was probably a missing incentive. Such performance regimes, recommended if not enforced by European legislation, were first experimented in Great Britain, and enforce compensations to be paid to the operator(s) by the infrastructure managers in case of traffic disruption caused by the latter. Such performance regimes are a necessary ingredient for franchising (competition for the tracks). In France, performance regimes were initially not envisaged, as they would have mostly resulted in a zero-sum game (transaction costs put aside) between the state-owned infrastructure manager and the state-owned, main operator. The other role of the performance regime, i.e. an incentive for improvement rather than a guarantee against low performance, was initially not considered.

With the 4th railway package, one would however expect performance regimes to become the norm. This is not yet the case, as a multi annual contract between SNCF Réseau and the State was signed in April 2017 for 2017-2026, without definition of performance targets to achieve.

Financial pressure

The French railway system is under pressure coming from budgetary restrictions, track access charges limits, accumulated under-investment (esp. renewals) and debt burden. The current French government wants to take the lead for the mobility in the future and started in September 2017 a think tank called ‘Assises de la mobilité ’to define future rail strategies.

6.5 Summary and conclusion

Innovation, overall, functioned well in the French case, with network capacity and performance remaining high. The benefits of innovation are however quite different depending on the market segment, with high speed leading and freight trailing. Partnership with Railenium opened the way to work with other partners such as industry and universities. Partnerships with start-ups allow fast development from idea to feasibility with POCs (proof of concept). As an example, TRAXENS and SNCF Logistics developed a project[16] called ‘Digital Freight Train’ which connects the train and the cloud, giving position of cargo and rolling stock, also allowing optimized train maintenance with sensors. As another example, INTESENS[17] developed connected solutions for the railway

[16] https://www.traxens.com/en/

[17] https://www.intesens.com/

maintenance adding sensors to monitor rail temperature, cumulative weight, OCL cable tension or lift monitoring, and sending notifications in real time. Digitalization seems to be on its way at a reasonable pace.

7. Slovenia

7.1 Introduction

For the purpose of the research on the incentives for introducing innovation in the rail system within the NeTIRail-INFRA project, interviews were conducted with three organizations related to the development of the railway infrastructure. For each organization, the scope of action is briefly presented. Interviewed organizations:

- Ministry of Infrastructure, Infrastructure Directorate, Sector for Railways (DRSI)
- Slovenian Railways – Infrastructure (SZ-Infra)
- DRI Investment Management (DRI)

7.2 Overview of the industry structure

The vertical hierarchy of railway system is as follows.

Government	Ministry of Infrastructure	Directorate for Infrastructure	Sector for Railways
	Slovenia State Holding	Slovenian Railways	SŽ-Infrastruktura
		DRI Investment Management	

A more detailed structure is shown overleaf. The key organisations are shaded in yellow.

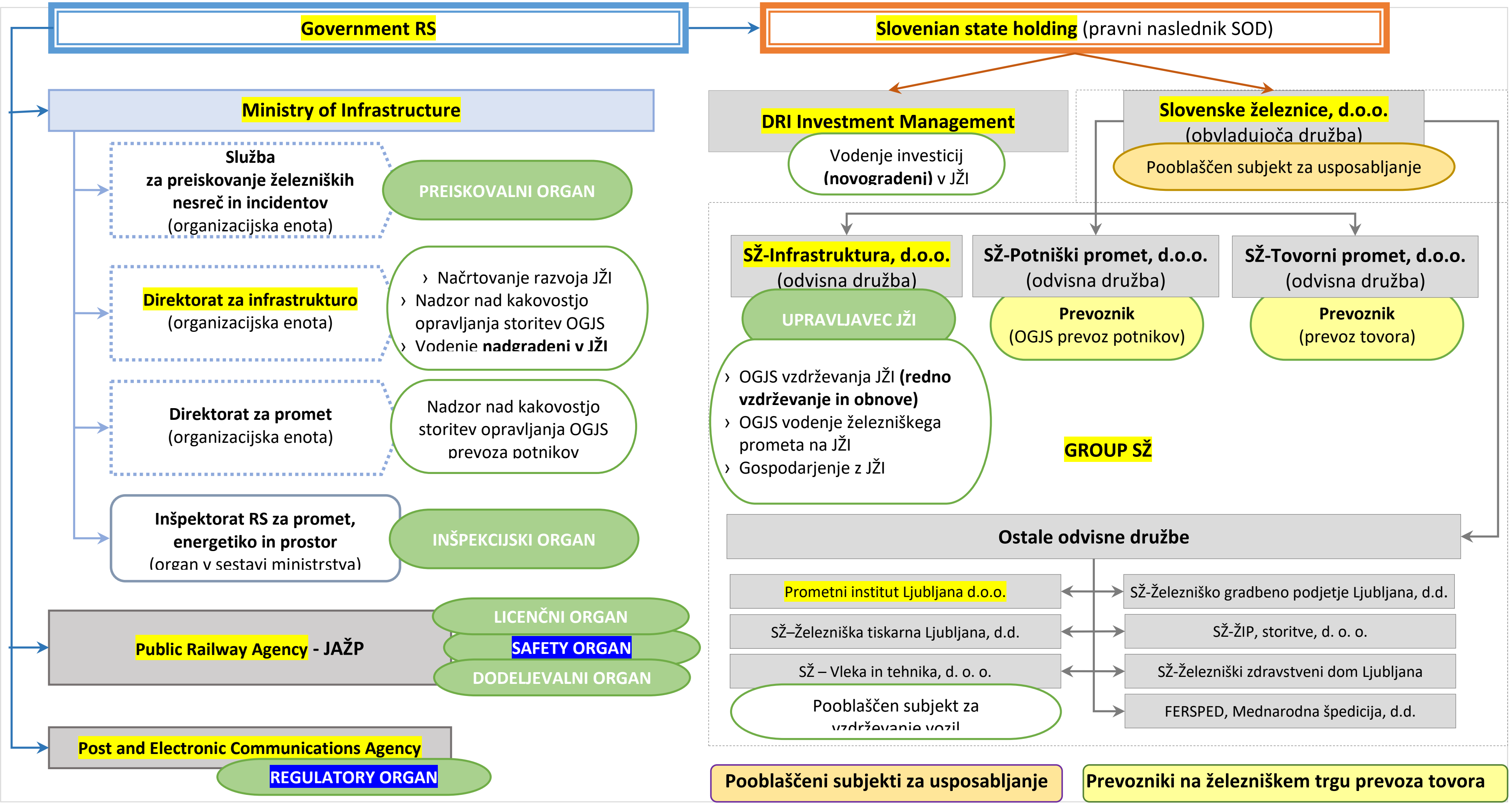
There is no institutional separation of the infrastructure manager from the carriers in Slovenia. The organization of the railways was followed by the form of a holding-integrated structure and a single company Slovenian Railway (Slovenske železnice d.o.o.) was transformed into the SŽ group. In the framework of this structure, the PRI manager and both national carriers are established as independent legal entities. SŽ-Infrastruktura, d.o.o. therefore, as a PRI manager, it cannot perform tasks related to the marketing of PRI and the determination and collection of user charges (access charges) for its use. These tasks are the responsibility of the Public Railway Agency. Company SŽ-Infrastruktura, d.o.o. prepares and publishes the network program and participates in the preparation and control of the implementation of the network timetable.

In Slovenia there is an open market in freight and passenger transport. On the basis of the valid national regulations, competition in the rail freight market is already in place in Slovenia and is established beside to the national carrier SŽ-Tovorni promet, d.o.o. For the provision of these transport services on PRI, a valid license and a safety certificate have been acquired by SŽ-Tovorni promet, d.o.o. and three other carriers, while no new carrier beside SŽ-potniški promet, d.o.o. has appeared in the railway passenger transport market despite the openness of this market. Passenger transport services are thus carried out as a compulsory commercial public service (OGJS) in rail transport, which is carried out by a national carrier SŽ-Passenger transport, d.o.o. according to a contract concluded by 2019.

The efficiency regime has been introduced in Slovenia on the basis of the Railway Transport Act and is further defined in the Network Program, published by the operator of the PRI - SŽ-Infrastruktura, d.o.o. Within the efficiency regime, the accuracy of the train is considered in relation to the allocated train path and is expressed as delays. Premature driving is not taken into account in the way in which efficiency is ensured. When calculating the compensation for the delay, only the delay arising from the primary causes created on the PRI network is taken into account. The delay compensation is calculated in such a way that, in relation to the responsible agent, the calculated delay is multiplied by a minute's delay of 0.10 EUR / min.

Countries have a Performance regime in line with Directive 2012/34 / EU (Article 35 / I), which sets out the basic legal framework for determining user charges (access charges) in the EU Member States, and Member States had to implement it by 16 June 2015at the latest. The efficiency regime is part of the charging scheme, by which rail operators and infrastructure managers are encouraged to minimize disorder in the use of the infrastructure and improve its use. Within the efficiency regime, all countries have established a system of incentives or penalties for delays.

Organization of Railway sector


Government RS
Slovenian state holding (pravni naslednik SOD)
Ministry of Infrastructure
DRI Investment Management
Vodenje investicij (novograđeni) v JŽI
Slovenske železnice, d.o.o. (obvladujoča družba)
Pooblaščen subjekt za usposabljanje
Služba za preiskovanje železniških nesreč in incidentov (organizacijska enota)
PREISKOVALNI ORGAN
Direktorat za infrastrukturo (organizacijska enota)
› Načrtovanje razvoja JŽI
› Nadzor nad kakovostjo opravljanja storitev OGJS
› Vodenje nadgradenj v JŽI
Direktorat za promet (organizacijska enota)
Nadzor nad kakovostjo storitev opravljanja OGJS prevoza potnikov
Inšpektorat RS za promet, energetiko in prostor (organ v sestavi ministrstva)
INŠPEKCIJSKI ORGAN
Public Railway Agency - JAŽP
LICENČNI ORGAN
SAFETY ORGAN
DODELJEVALNI ORGAN
Post and Electronic Communications Agency
REGULATORY ORGAN
SŽ-Infrastruktura, d.o.o. (odvisna družba)
UPRAVLJAVEC JŽI
› OGJS vzdrževanja JŽI (redno vzdrževanje in obnove)
› OGJS vodenje železniškega prometa na JŽI
› Gospodarjenje z JŽI
SŽ-Potniški promet, d.o.o. (odvisna družba)
Prevoznik (OGJS prevoz potnikov)
SŽ-Tovorni promet, d.o.o. (odvisna družba)
Prevoznik (prevoz tovora)
GROUP SŽ
Ostale odvisne družbe
Prometni institut Ljubljana d.o.o.
SŽ–Železniška tiskarna Ljubljana, d.d.
SŽ – Vleka in tehnika, d. o. o.
Pooblaščen subjekt za vzdrževanje vozil
SŽ-Železniško gradbeno podjetje Ljubljana, d.d.
SŽ-ŽIP, storitve, d. o. o.
SŽ-Železniški zdravstveni dom Ljubljana
FERSPED, Mednarodna špedicija, d.d.
Pooblaščeni subjekti za usposabljanje
Prevozniki na železniškem trgu prevoza tovora

Agreements are set up by the operator and carriers to be a joint body to deal with the disputed cases. If one of the parties does not agree with the distribution of delays, it may appeal to the regulatory body.

Multi annual contracts also define the achievement of a more cost-effective form of provision of the services of the operator, which will be adapted to the needs of the user. SZ-Infrastructure has a multi-annual maintenance contract with the Ministry of Infrastructure (2016-2020), supplemented with annexes (if necessary). Slovenian Railways conclude contracts with each operator (driver/transporter) for access or transport by public rail infrastructure. The implementation mode or performance regimes are defined prescribed in the annual Network Program.

The operator must strive to increase the efficiency and effectiveness of services by introducing new technologies. In the interests of safety, maintenance and improvement of the quality of infrastructure services, the manager (operator) shall be provided with incentives to reduce the costs of maintenance and provide the infrastructure that affect the level of user charges (track access charges).

On the basis of the Railway Transport Act, the Railway Efficiency Regime is established, which aims to encourage carriers and operators to reduce disruptions in the rail network and improve the quality of the performance of the transport service (Directive 2012/34 / EU).

7.3 Ministry of Infrastructure, Infrastructure Directorate, Sector for Railways

DRSI is a body within the Ministry of Infrastructure that prepares calls for the execution of works, including the preparation of a project task with technical specifications, based on the valid regulations signed by the Ministry of Infrastructure.

Innovation can be new (more modern) elements of the infrastructure or processes or approaches, expressed in the technology of work execution, if policies allow. In this case there are no specific procedures for introducing innovations. Innovation depends on the market.

The incentives, which are mainly in interest of producers, are the reduction of labour costs, costs of materials, etc. and thus greater competitiveness on the market. Incentives include lower maintenance costs, lower consumption, following environmental trends and / or requirements.

7.4 Slovenian Railways – Infrastructure

SZ-Infrastructure is a company in the group of Slovenian Railways, a public railway infrastructure manager, responsible for its maintenance, in accordance with national safety and other regulations. The system of regulation determines the implementation of budget spending. The company is involved in its operations with carriers, research institutions and universities.

Innovation should be encouraged and financed by manufacturers of infrastructure elements or equipment. Carriers and manufacturers have their own innovation systems and are doing their product updates. SZ-Infrastructure is in the role of a buyer on the market.

Incentives for possible innovations within the system of Slovenian Railways are defined in the "Instruction on the Inventions from the Employment of the Company Slovenian Railways" signed in 2010. The innovations proposed by the employees of the Group are promoted with cash prizes based on the evaluation of the invention. Incentives can be at different levels, not just financial; incentives for introducing innovations are also easier maintenance of infrastructure, less necessary labour, various adjustments / simplifications of work, etc.

7.5 DRI Investment Management

DRI Investment Management is the Company for the Development of Infrastructure, owned by the Republic of Slovenia, established with the intention to provide, as an internal operator, public utility services, investment engineering services, investments in public infrastructure and other advisory services. The role of the company is the support of the Ministry of Infrastructure.

A wide range of company activities enables organization of the management of the entire investment: from preparation, design, construction to maintenance and management of all types of infrastructure facilities. A comprehensive range of consulting and engineering services is rounded off with specialized consulting and research services.

The company is faced with the introduction, implementation of innovations in managing projects of new buildings and upgrades. DRI offers the most competent assistance in order to help the Ministry of Infrastructure to perform changes in the field of new railway infrastructure construction.

7.6 Interviews on Incentives for the Introduction of Innovations in the Railway System

The following three key issues were asked to the organizations:

a. What are the main difficulties that either prevent or slow down the introduction of innovations in the railway system?

b. How could it be made easier for innovations in the rail system to be introduced?

c. Do the different organisations in the rail sector have appropriate incentives to develop and introduce innovation?

In the context of the interviews, the following general issues were also included:

d. What incentives for railway innovations exist in Slovenia?

e. How to prepare proposals for railway innovation?

f. What is the role of the ministry in DRSI / SZ-Infra / DRI?

The responses of individual organizations are united for the key three issues/questions of introducing innovations in the rail system.

a. **The main difficulties that prevent or slow down the introduction of innovation**

- the presence of a small number of Slovenian companies, the industry of the railway sector, producers of railway elements at international meetings, therefore there is (almost) no Slovenian development and it is more or less to follow the European or the global market;

- the rigidity of the rail system, the long-term nature of the introduction of changes, the development in the railway sector is slower; the process of implementing innovation is complex

- a high threshold for entering the market through demanding verification procedures and high investment costs, a large financial investment and the input of staff is required;

- a lack of experience from the past, the introduction of innovation is difficult

- the most important administrative barriers - national regulations, by-laws that do not follow market innovations, the change of regulations is not flexible, with domestic and international regulations due to the need to ensure interoperability between the railway networks

- the ability to provide safe, reliable, highly available systems

- lack of expertise and connection between institutions

b. **Incentives for introducing innovations**

- State stimulation or the competent state body that monitors market innovations abroad, and follows them in such a way as to prepare a tender that corresponds to the trend of novelty, change and no longer allows the old one;

- modification of the presented innovations at fairs, taking into account as far as legislation permits
-SZ-Infrastructure cooperates with external initiators and helps them in innovation: it enables testing and obtaining permits for use, certificates of acceptability, suitability in terms of existing infrastructure and vehicles

-test polygon for verifying various solutions, with simple, equivalent access for all organizations or individuals, and support from EU funds

- EU-wide regulation with minimal impact of local / state institutions (TSI leaves too vague space)
- management of solutions in larger corporations, organization of various incentives for employees in enterprises

- financial incentives by the EU resources

c. **Adequacy of incentives**

- there are other institutions with options and with their incentive mechanisms (the Ministry of the Economy)

- the incentives are not the most appropriate, in addition to the invention instructions (for SZ), there are no specific other dedicated resources for innovation

- no right direction on how the railway can cooperate with manufacturers and suppliers

-with TSI, the scope of regulation at EU level began to decrease, but there are no corresponding country-specific regulations, infrastructure managers cannot easily check and innovate – because of focus on efficiency of operations, reduction of costs and staff numbers, due to software complexity of the solution, the lack of knowledge, the lack of connections between organizations, the unevenness of ownership, management rights
* *There are certain types of incentives that are not the right incentive for innovation: the payment of user charges for train paths or transport by infrastructure; it is an incentive for PRI carriers and managers to reduce disruptions in the rail network and improve the quality of the performance of the transport service so as not to cause delays, otherwise compensation will be charged for the occurrence of delay.*

The Ministry of Economic Development and Technology, which has established better mechanisms for stimulating innovation, is involved in innovation more than the railway organization. In the field of technology development, a sub-area "Promoting Innovation and Technological Development" has been established, which generally refers to entrepreneurship and not exclusively to the rail sector.

7.7 Promoting Innovation and Technological Development

In order to create an innovative economy, investment in research and technological development is one of the key factors for the competitive ability of companies.

The aim of implementing measures is to increase entrepreneurial investments in research and development, to promote the integration of the Slovenian economy into international scientific research programs, to promote employment or to train researchers and developers in the economy, and to build an innovation infrastructure in support of the national innovation system.

Technological development measures – acceleration of private R & D investment - include repayable and non-refundable funds to promote R & D activities in companies and R & D tax incentives.

Repayable and non-refundable funds are called by the following institutions:

SPIRIT SLOVENIJA JAVNA AGENCIJA

SLOVENSKI PODJETNIŠKI SKLAD

SID Banka

Companies can also apply tax incentives for investments in research and development. Legislation governing tax relief:

“The Corporate Income Tax Act” (for companies) and the “Personal Income Tax Act” (for sole proprietors) stimulate that taxpayers can reduce their tax base by 100% of their own investments in research and development (R & D). The condition is that the taxpayers have a tax base, but the possibilities of planning in the case of tax relief are usually better than in the case of subsidies.

International Development Cooperation is an excellent source for access to the latest knowledge and integration into the most advanced networks and consortia in the field of technological development.

- EUREKA
- EUROSTARS
- Research Fund for Coal and Steel
- Cooperation with the European Space Agency – ESA
- HISTORY 2020
- ECO INNOVATION

Legislation or regulation in the field of railway innovations: the field of introduction of novelties in the railway system is regulated by the "Railway Traffic Safety Act" and on its basis, by-laws, including the following Regulations: "Rules on the Upper Structure of Railways", "Rules on the Lower Structure of Railways", "Rules on Railway signalling devices "," Rules on the railway telecommunication network".

In addition to the law and implementing regulations, the safety requirements prescribed in many safety standards (TSI, ISO and UIC) must be observed.

Within the Slovenian Railways system, only the "Instruction on the Inventions from the Employment Relationship of the Company Slovenian Railways" was adopted on innovations.

As noted above, there is a holding company structure adopted, and multi-annual agreements are in place aimed at improving efficiency in part through technological developments.

8. Conclusions

The NeTIRail-INFRA project is concerned with innovation in the rail sector, and particularly in rail infrastructure. This report particularly concerns the incentives to innovate in the European rail sector, where under EU law infrastructure is separated from operations (at last as separate subsidiaries of a state owned holding company, but often completely separated), there is open access for new freight and international passenger operators (and increasingly for domestic passenger, either through competitive tendering for public service contracts or on a purely commercial basis, and there is a regulator responsible for regulating track access charges and ensuring non-discrimination.

In this report, we have examined documentary evidence and undertaken interviews in Britain, Sweden, Germany, France and Slovenia. By far the most thorough and extensive review is that undertaken for Britain, and Britain is the country that has paid the most attention to incentives, through sophisticated track access charges and performance regimes, through various schemes ranging from sharing benefits of efficiency gains on a particular route to deep alliances sharing all costs and revenues for a particular train operator, and through many specific innovation funds. In many other cases there is no attempt to design schemes to align incentives. For instance, neither France nor Germany has a performance regime, and in both countries track access charges are levied per train kilometre, with no attempt to differentiate according to gross weight and other aspects of track friendliness.

In general, it was concluded that major reasons for a lack of incentives to innovate derived primarily from two sources:

1. Fragmentation of the industry, with the result that the organisation undertaking interviews may not be the same as the one receiving the benefits, alongside inadequate use of mechanisms such as appropriate track access charges to ensure appropriate incentives exist. Highly differentiated track access charges, which ensure that train operators receive the benefits of innovations that reduce wear and tear on the track or improve capacity utilisation will reduce the problem, as will sophisticated performance regimes which ensure that both infrastructure manager and train operators have incentives to improve performance; in the case of the infrastructure manager such incentives may be expected to continue indefinitely. But it remains the case that train operators have little incentive to help the infrastructure undertake infrastructure work efficiently (for instance by rescheduling services to allow longer possessions).

2. A short term emphasis in regulatory and franchising arrangements, giving too little incentive to achieve long run benefits. Of course, the infrastructure manager does have long time horizons, so this is particularly a problem with relatively short franchises. For instance, when choosing rolling stock, the incentive on franchisees is to select rolling stock that will enter service quickly and reliably rather than considering innovations which may only have longer term benefits,

But regulatory systems that place particular emphasis on targets for limited control periods or multi annual contracts may distract from the longer term, whilst binding cash limits which apply even to cost-reducing investments may make it impossible to consider longer term impacts if they add to current costs. On the other hand, a failure to use such regulatory instruments may reduce pressure to cut costs in what remains a monopoly supplier of infrastructure. Another way of reducing the degree of monopoly is by contracting out functions of the infrastructure manager such as maintenance, renewals and new projects, leaving essentially only the functions of the systems operator as a monopoly. But again there is a risk that the contracts may be relatively short and may be written in a way which gives limited scope for innovation.

Both effects may be mitigated by specific arrangements in one or more of the countries examined, but these arrangements themselves have disadvantages as well:

a. The presence of a holding company in Germany which plays an active role in ensuring that activities by one part of the group are appraised in terms of the impact on the group as a whole, and which is very much concerned with the long term future of the business, including promoting innovation. But the presence of such a holding company has been seen as a potential barrier to competition; at the least it makes the task of the regulator in ensuring no discrimination more difficult, and the larger the share of the market taken by competitors, the less effective the holding company will be at taking a comprehensive view of the industry.
b. Arrangements which ensure that train operators bear all of the costs of the infrastructure manager in the form of track access charges, and thus have an incentive to work together to reduce them. But this will certainly add to the risks of train operators, potentially leading to less competition and higher subsidy requirements in bids for franchises. Another problem here is that to give the correct incentives in terms of levels of service track access charges need to be based on marginal cost and econometric evidence suggests that – unless there are high congestion or scarcity costs – marginal cost is well below average cost. Even if some costs of capacity can be charged to the train operator as a fixed charge based on the avoidable costs of their services, the presence of joint costs means that this may not exhaust the total costs of the infrastructure manager. Allocating these joint costs to a specific train operator (for instance as the 'prime user' of the facilities in question) may lead to attempts to free ride, for instance by reducing services to the point where they are not the prime user.
c. Cost and revenue sharing arrangements between the infrastructure manager and the train operator.

Obviously in a sense making infrastructure and operations part of the same holding company is the ultimate way of sharing costs and revenues provided that mechanisms exist to ensure that subsidiaries do take impacts on the company as a whole into account, rather than just impacts on themselves. A less extreme approach has been tried in Britain, in which infrastructure manager and train operating company make specific agreements to share costs and revenues, through deep alliances or less ambitiously through the Route Based Efficiency Benefit sharing mechanism, although the evidence is that achieving such agreements on a voluntary basis is difficult – train operators do not want to take on infrastructure costs risks and infrastructure managers do not want revenue risk. A way of overcoming this would be to let vertically integrated franchises, in which the train operator took control of the infrastructure for the duration of the franchise. Provided that the franchise was long, this offers in principle the possibility of overcoming both fragmentation and short time horizons. But as always there are disadvantages. Long franchises themselves reduce the level of competition, and the arrangement will only have the desired effect if most services in the area in question are part of the same franchise. This may conflict with promotion of freight or open access passenger operations, and with the design of efficiently sized franchises focussed on specific market segments.

In short there is no perfect solution to the problem of incentives, and the best course of action will vary with circumstances. For instance, if the infrastructure in a particular area is almost exclusively used for subsidised passenger services then a vertically integrated franchise will be more appropriate then if there is also a lot of commercial passenger and/or freight traffic. Where it is considered desirable that a large share of train operations will remain in the public sector, then a holding company solution may make sense, but it must be borne in mind that new entrants to the industry may bring a new and innovative approach, so blocking new entry may not promote innovation.

In short, there are no simple answers, but the issue of incentives needs much more consideration in designing rail reforms than it has often had, in order to give the best possible incentives in the situation under consideration.

9. References

Arafer's website: http://www.arafer.fr/le-ferroviaire/tarification-infrastructure-ferroviaire/la-definition-du-cout-directement-imputable/

Brown, R. 2013. *The Brown Review of the Rail Franchising Programme*. London: Department For Transport.

Caves, M., Doyle, C., 2007b. *Network Separation and Investment Incentives in Telecommunications*, August, ThinkTel, Milan, Italy.

Department For Transport (DFT). 2016. *Rail Freight Strategy*. London: Department For Transport.

Doherty, A. 2016. Shift2Rail has the potential to transform infrastructure. International Railway Journal. (online). (Accessed on 28 November 2017). Available from http://www.railjournal.com/index.php/track/shift2rail-has-the-potential-to-transform-infrastructure.html

Ekman, J., A. Holst, M. Aronsson, M. Forsgren, M. Bohlin, S. Larsen. 2006. *TIME – en gemensam informationsutbytesplattform för järnvägstransportbranschen*. SICS Technical Report T2006:03.

EPSF's website: http://www.securite-ferroviaire.fr/

Espling, U. 2007. *Maintenance Strategy for a Railway Infrastructure in a Regulated Environment*. Doctoral Thesis, Luleå University of Technology, Division of operation and maintenance engineering, Luleå Railway Research Center.

European Commission. 2014. Shift2Rail Strategic Masterplan. Brussels: EC. (online). (Accessed on 28 November 2017). Available from https://ec.europa.eu/transport/sites/transport/files/modes/rail/doc/2014-09-24-draft-shift2rail-master-plan.pdf

Council Directive 2012/34/EU of the European Parliament and of the Council. Of 21 November 2012. *Establishing a Single European Railway Area* (Recast).

HackTrain. 2016. *The BARRIERS report: Bringing Actionable Recommendations to Revitalise Innovation and Entrepreneurship in the Rail Sector*. London: Innovate UK.

Holst, A., M. Bohlin, J. Ekman, O. Sellin, B. Lindström, S. Larsen. 2012. Statistical Anomaly Detection for Train Fleets. *Proceedings of the National Conference on Artifitial Intelligence*, 3, pp 2217-2223.

INTESENS : https://www.intesens.com/

Jack, A. 2017. Research UK: Looking to a Bright Future. *Railway Gazzette International*, April, pp 46-48.

Karlsson, P., L. Redtzer. 2000. *Utvecklingsstrategi för Banverkets produktionsverksamhet: Förutsättningar och strategier för konkurrensutsättning av Banverkets produktionsverksamhet*. GD00-2827/01.

Klein, B., Crawford, R. and Alchian, A. 1978. Vertical Integration, Appropriable Rents, and the Competitive Contracting Process. *Journal of Law and Economics*, 21 (2), pp 297-326.

McNulty, R. 2011. *Realising the Potential of Rail in Great Britain- Final Independent Report of the Rail Value for Money Study*. London: Department For Transport.

Merkert, R., Smith, A. and Nash, C. A. 2012. The Measurement of Transactions Costs - Evidence from European Railways. *Journal of Transport Economics and Policy*, 46 (3), pp 349-365.

Mizutani, F., Shuji Uranishi. 2013. Does Vertical Separation Reduce Cost? An Empirical Analysis of the Rail Industry in European and East Asian OECD Countries. *Journal of Regulatory Economics*, 43 (1), pp. 31–59.

Mizutani, F., Smith, A.S.J., Nash, C.A. and Uranishi, S. 2015. Comparing the Costs of Vertical Separation, Integration, and Intermediate Organisational Structures in European and East Asian Railways, *Journal of Transport Economics and Policy*, 49 (3) July, pp. 496-515.

Multi annual contract https://www.sncf- reseau.fr/sites/default/files/upload/_Mediatheque/rapports-annuels/Contrat_pluriannuel_Etat_SNCF_Reseau-2017.pdf

Nash, C. A., Smith, A., van de Velde, D., Mizutani, F. Uranishi, S. 2014. Structural reforms in the railways: Incentive misalignment and cost implications, *Research in Transportation Economics*, 48, pp.16-23.

Nilsson, J-E. 2016. *Kvalitetsavgifter – Problem och tänkbara lösningar*. VTI-rapport 2016:884

Odolinski, K. 2015. *Reforming a publicly owned monopoly: costs and incentives in railway maintenance*. Doctoral Dissertation, Örebro Studies in Economics 30.

Odolinski, K. 2016. Contract design and performance of railway maintenance: effects of incentive intensity and performance incentive schemes. *CTS Working paper* 2016:20, Centre for Transport Studies, Stockholm.

Odolinski, K. and Smith, A.S.J. (2016). Assessing the cost impact of competitive tendering in rail infrastructure maintenance services: evidence from the Swedish reforms (1999-2011). *Journal of Transport Economics and Policy*. 50(1), 93-112.

Office of Road and Rail (ORR). 2013. *Final Determination of Network Rail's Outputs and Funding for 2014-19*. London: Office for Road and Rail.

Organisation for Economic Cooperation and Development (OECD). 2001. *Recommendation of the Council concerning structural separation in regulated industries*. Paris: OECD.

Organisation for Economic Cooperation and Development (OECD). 2003. *The benefits and Costs of Structural Separation in the Local Loop*, DSTI/ICCP/TISP (2002)13/FINAL. Paris: OECD.

Organisation for Economic Cooperation and Development (OECD). 2005. *Structural Reform in the Railway Industry*, DAF/COMP (2005) 46. Paris: OECD.

Partenariat d'innovation: https://www.economie.gouv.fr/daj/partenariat-innovation-2016http://www.alstom.com/fr/press-centre-francais/2016/9/sncf-et-alstom-lancent-leur-1er-partenariat-dinnovation-pour-creer-la-nouvelle-generation-de-tgv/

Pittman, R. 2007. Options for restructuring the state-owned monopoly railway. *Research in Transportation Economics*, 20, pp. 179–198.

Rail Delivery Group (RDG). 2016. *Review of Charges*. London: RDG.

Railway Safety and Standards Board (RSSB). 2017. *Inspiring and Supporting Railway Innovation in the UK*. London: RSSB.

Raimbault, N.; Banquart, C.; Poinsot, P. 2017. *Innovations in the railway sector: an innovation system in transition between state impulsion regime and market oriented regime*, ISTE open science London.

Reforme marchés publics : https://www.economie.gouv.fr/entreprises/marche-public-reforme

Shaw, N. 2016. *Shaw Report: The Future Shape and Financing of Network Rail*. London: Department For Transport.

SNCF websites: https://www.sncf-reseau.fr/fr/a-propos/presentation http://www.sncf.com/ressources/reports/rapport_financier_annuel_groupe_sncf_2016.pdf

Sveriges Riksdag. Förordning (2008:1300) med instruktion för Transportstyrelsen, Svensk författningssamling 2008:1300, 3 §, task 3), https://www.riksdagen.se/sv/dokument-lagar/dokument/svensk-forfattningssamling/forordning-20081300-med-instruktion-for_sfs-2008-1300

TRAXENS : https://www.traxens.com/en/

Funded under the H2020 programme

Collaborative project H2020-MG-2015-2015 GA-636237

Needs Tailored Interoperable Railway – NeTIRail-INFRA

Deliverable D1.7
Incentives Final Report Annex 2
The Impact of Quality on Cost
Is higher quality costly? Marginal costs of quality: theory and application to railway operation

Document ID: NeTIRail-WP1-D1.7v1.0-FINAL – ANNEX2

Due date of Deliverable: 30/09/2017

Actual submission date: 21/12/2017

Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Task leader for this deliverable: Professor Andrew Smith, Institute for Transport Studies, University of Leeds

Document status		
Revision	Date	Description
V0.1		First draft (authors Andrew Smith, Manuel Ojeda Cabral)
V0.2	14/12/2017	Review by USFD and ALU-FR
V1.0	21/12/2017	Final version
Reviewed	YES	

Executive Summary

Travel time reliability is a key element of any transport system. An element of quality. In the railway sector, much has been discussed about the costs of delays to passengers and their willingness-to-pay to reduce them, i.e. the demand side of the market. However, delays in the supply side of transport markets have received far less attention (Van Oort, 2016). Similarly, quality in railway cost studies has often been neglected or considered in an ad-hoc basis. This paper fills several gaps in the transport and railway literature by studying the relationship between the costs of railway supply and the degree of travel time reliability. First, we articulate a generic theoretical framework for the relationship between costs and quality. We introduce the notions of marginal proactive cost and marginal reactive cost, leading to a U-shaped relationship between cost and quality. The framework acknowledges that low reliability could be associated with higher (reactive) costs but, similarly, high reliability may need high (proactive) costs too. Secondly, we apply the framework to a dataset of train operating companies (TOCs) in the UK over a period of five years. The estimated cost model allows us to empirically observe the cost-reliability relationship and obtain estimates of the marginal costs of improving reliability. The framework and analysis can be used to aid quality related decisions of TOCs, Infrastructure Managers and regulators in the railway industry. The proposed framework can also be applied to other cost-quality contexts, in and outside transportation.

Table of contents

Abbreviations and acronyms

Abbreviation / Acronym	Description
PPM	Public Performance Measure
DfT	[UK] Department for Transport
IM	Infrastructure Manager
TOC	Train Operating Company
PAF	Prevention Appraisal Failure
COQ	Cost of Quality
MC	Marginal Cost
MCq	Marginal Costs of quality
MPCq	Marginal Proactive Cost of quality
MRCq	Marginal Reactive Cost of quality
FE	Fixed Effects
RE	Random Effects

1. Introduction and background

The reliability of a transport network is an important element of the market, both from the supply and from the demand perspective. *Reliability*, here understood as the degree of certainty of travel times[1], has been mainly studied from the point of view of travel demand. Unreliability is costly and unpleasant to transport users and hence affect travel demand negatively. An extensive body of literature has studied how passengers value reliability (for a review, see Carrion and Levinson, 2012) and how changes in reliability affect demand (see Wardman and Batley, 2014). However, the relationship between reliability and the supply side of the market has surprisingly received little attention. A limited number of papers and reports have been identified in this area and none of them provide a theoretical discussion.

It is therefore necessary to increase our understanding of the cost-reliability relationship. Nonetheless, we shall see that this relationship is not straightforward. Reliability is an element of quality of a transport system. In regulatory contexts, it is generally assumed that if firms or regulators wish to improve quality, these improvements should come at a cost (i.e. positive marginal cost of quality). This would translate into finding a positive direct relationship between cost and quality under standard cost studies. However, there is evidence finding a negative relationship, hence indicating that poor quality can be associated with higher costs. Examples of this negative relationship have been found in the water sector in England and Wales (2014 Periodic Review), in health economics (Gutacker et al. 2013) and in the airlines sector in the US (NEXTOR, 2010). A negative relationship between cost and quality can create perverse incentives for economic regulation.

Secondly, conceptually and methodologically it is also important to pay more attention to quality in the context of cost, productivity and efficiency studies. The vast literature in this area often neglects quality, possibly because it is hard to obtain reliable and comparable data. But perhaps also because theoretically the relationship has not been formalized and hence there is a lack of clarity. Yet, quality is undoubtedly a significant factor of the costs of a firm. By extension, we can also expect a relationship between quality and efficiency or productivity. The link between costs and quality seems, in any case, the first step.

1.1 The UK railway context

In the railway sector in the United Kingdom, train delays and cancellations are 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) and customers' satisfaction surveys indicate there is still a long way to go (e.g. House of Commons Transport Committee, 2017). The figure below shows how train punctuality, measured by PPM, has stagnated over the last decade around 89% on-time levels[2]. Even more worrying is that industry reports have

[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 notion of reliability of travel times.

[2] This 'on-time' measure, known as the Public Performance Measure (PPM), includes allowances on top of the scheduled travel time: a train is 'on-time' if it arrives at its final destination within 5 minutes of the scheduled arrival time, or within 10 minutes for Long Distance services.

shown that passengers do not perceive this measurement (PPM) to be a fair reflection of the industry under-performance (House of Commons Transport Committee, 2017).

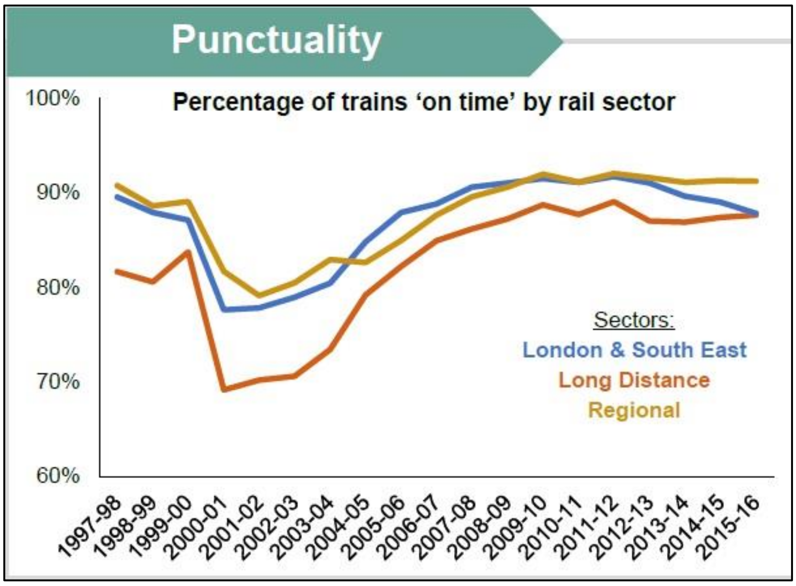


Figure 1. Rail punctuality in the UK. Source: DfT (2017), Rail Trends Factsheet

While there is evidence of how much travellers are willing to pay for reliability improvements in the UK (e.g. ARUP et al., 2015), we do not really know how much (if anything) does it cost to improve reliability from the supply perspective. The British railway is a vertically separated industry, like most European countries. Rail track management is the responsibility of the Infrastructure Manager (IM) and operation is conducted by Train Operating Companies (TOCs). A regulator, namely 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 IM and TOCs. In Britain, the performance incentives system is called Schedule 8 (S8):

> *Schedule 8 is an automatic mechanism for ensuring that train operators and Network Rail are held financially harmless for delays that they cause to each other. The ORR sets the targets and the rates. A formula drives the payments, based upon who caused the delay, how bad the delay was, and how much fare box revenue is estimated to have been lost (now and in the future) from the incident. If everyone achieves target level performance, no money changes hands* (Network Rail, 2016).

The incentives system 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 the target agreed in the contract. However, how costly is it for the IM and TOCs to improve reliability? An answer to this questions would seem necessary for the design of any performance incentives system. Yet, it is far from clear whether we know the answer.

1.2 Research objectives

This papers builds upon the existing railway and transport literature, while drawing ideas from other bodies of literature to fill the existing gaps. The contribution of this paper is two-fold. First, the gaps in the literature make it necessary to enhance our understanding and formalize the relationship between cost and quality. More precisely, we address the question: 'what is the relationship between the costs of railway supply and the reliability of services'? We develop a generic framework that provides a theoretical foundation for cost elasticities with respect to quality and marginal costs of quality. Secondly, through the lenses of this framework, we empirically address the relationship between the cost of TOCs and the reliability of their services. 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 Manager (IM), other transport modes and, more broadly, other areas where quality is relevant. In particular, the context of the IM is the focus of a different paper (Gillies-Smith et al., forthcoming).

The remainder of the paper is structured as follows. In the next section we present the literature review. Section 3 contains our theoretical framework for the cost-quality relationship and its application to railway reliability. Section 4 introduces the methodology used for the empirical work, followed by a description of the dataset. The results are presented and discussed in section 5. Section 6 concludes.

2. Literature review

2.1 Railway context

Affuso et al. (2002) measured the efficiency of British TOCs soon after privatisation took place, and tried to relate 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 economic efficiency. However, they acknowledged that the link was statistically very weak. Kennedy and Smith (2004), in a study of the efficiency of the IM in the UK (Network Rail), found a weakly significant negative relationship between delays (caused by the IM) and total costs. They also report a negative relationship of the total costs with the number of broken rail incidents, a factor that may be linked to both safety and reliability. Abate et al. (2013) interpret those findings as evidence that improving reliability is costly, something they label as the "effort effect".

Abate et al. (2013) conducted a European study of the relationship between reliability and productivity in the railway. Their paper crucially recognises the presence of countervailing forces when studying reliability. They describe up to 3 mechanisms that link punctuality and productivity, which they refer to as "effort effect", "utilization effect" and "demand effect". Their view is that the overall relationship is ambiguous due to the presence of these countervailing forces (the effort effects implies a negative relationship between punctuality and productivity, while the other two effects imply the opposite). However, they do not formalize the relationship and directly resort to empirical analysis to try to answer their questions. Applying Data Envelopment Analysis (DEA) and a Malmquist index to data from seven European railways, they conclude that improving reliability is not necessarily linked to reductions in productivity. They interpret the empirical results as evidence that the three mentioned effects compensate each other in practice.

On a somewhat more distant but still related application, Estache et al. (2007) dealt specifically with the quantity-quality trade-off in the Brazilian freight railway. Using a Mamlquist productivity index, they find evidence of the existence of such trade-off in some periods but also evidence that quality can be positively correlated with quantity under a new different regulation scheme. 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 improvements of reliability 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 costs for a transport operator, but does not enter into more details.

Overall, the scarce existing evidence in the railway literature provides only a fairly basic discussion on the link between railway supply costs (or efficiency or productivity) and the degree of reliability of their services. A basic theoretical articulation of the expected relationship (at least between costs and reliability) is missing. Such framework, at least conceptually, seems necessary prior to any empirical work.

Additionally, there are several issues with the existing empirical works. For example, Abate et al. (2013) and Affuso et al. (2002) tried to relate reliability with productivity (a function of costs), and therefore did not explore the more primary link between reliability and (observable) costs. Furthermore, the data used for their work 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 some sort of aggregate index. Finally, Kennedy and Smith (2004) research questions are indeed closer to ours, but they focused on IM and the lack of theoretical underpinning undermined their analysis.

2.2 Other contexts: outside the railway

Reliability is also important in the air transport sector. NEXTOR (2010), in an extensive report, 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, and followed it with the estimation of a translog cost model that includes delay time as an explanatory variable of the costs. They find a significant positive effect of the combination of delay and buffer time on airline costs, and use the model results to derive some estimates of the total cost of delay time to airlines for a given year. A drawback of their approach is that 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. More recently, Peterson et al. (2013) carried out a new analysis of the overall cost of airline delays to the economy using a general equilibrium model. Such aggregated approach would not be suitable to answer our research questions. These works seem to be more concerned with ‘how costly are delays’, and less with ‘how costly are delay improvements’. Other air-transport studies focused on efficiency and performance (Assaf et al, 2014; Merkert et al., 2015, Link, 2017). Link (2017) also highlights the lack of research in quality from the supply perspective in the air transport literature, similar to our findings in the railway counterpart.

Since quality (not restricted to punctuality) is relevant in other sectors, we explored the literature outside transport. First, we explored the wider literature on quality and costs. 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 costs types, leading to what is known as the Prevention Appraisal Failure (PAF) model. They categorized the costs associated to 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 figure 2).

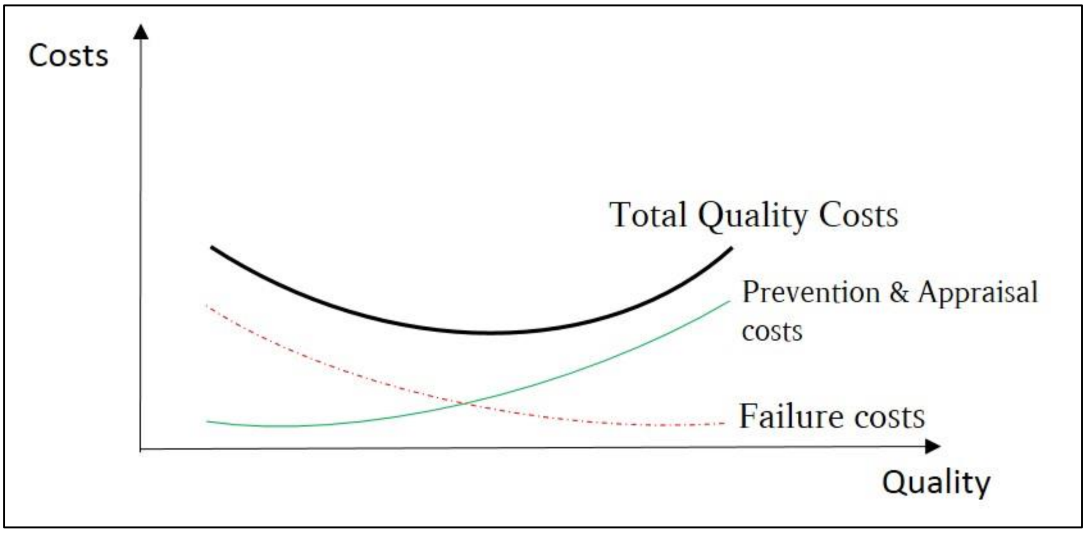


Figure 2. Illustration of a classical Cost of Quality Model

The COQ model depicted in figure 2 (PAF model) is regarded as the most popular and widely used model in this area (Schiffauerova and Thomson, 2006). This and other models that relate costs and quality have been developed since (e.g. Crosby, 1979, Juran and Gryna, 1988). For reviews on this body of 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, to the best of our knowledge, none of the theoretical and empirical works focused on the analysis of the economic concepts of cost elasticities and marginal costs. Instead, they have looked at total costs and used accounting, management and simulation methods to unpick the different types of quality costs.

Secondly, we also explored the literature from other utility sectors (e.g. health, water, energy) where practitioners could share interests similar to ours. Here we found more examples of theoretical frameworks for the relationship between cost and quality combined with empirical applications. Interestingly, these applications provide frameworks that can be seen as equivalent to the wider Cost of Quality Models (COQ), such as the PAF model. However, they all seem unaware of the existence of this more generic body of literature that focuses on quality and costs.

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. 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. However, again there is no link with generic COQ literature and no formalization of the model.

In summary, both the wider literature on quality 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. We hope to do so and develop a more comprehensive and generic framework for the relationship between cost and quality, where marginal costs are at the centre of the discussion. Our framework will be applied to the study of railway reliability, but will be general enough to be applicable within a wide range of quality-cost contexts.

3. Theoretical framework

This section presents the theoretical framework for the relationship between costs and quality. The reliability of train operating companies will be a specific case that fits into the generic framework.

3.1 A generic framework for the relationship between Quality and Costs

We develop a generic theoretical framework for the cost-quality relationship. The framework explicitly addresses the cost-quality relationship in terms of marginal costs. The framework can be derived from the classical COQ model known as PAF model. It also brings together the existing literature from different utility sectors (mainly transport, health and energy) with the wider economics and management literature on cost of quality models. In particular, it builds upon the frameworks developed by Juran et al. (1962) in quality research and Jamasb et al. (2012) in the electricity sector and Gutacker et al. (2013) in the health sector.

We first introduce the framework conceptually, a generalization of Jamasb et al. (2012)'s model. This is depicted in figure 3 below.

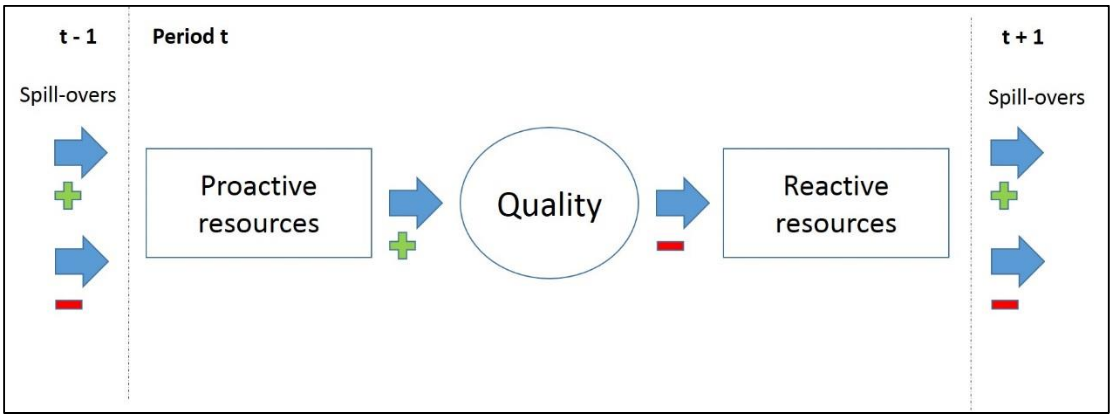


Figure 3. Conceptual Framework - Quality and use of resources

The key element of the relationship between Quality and Costs 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. 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. The term proactive more generally accommodates any use of resources aimed at improving quality. Proactive costs hold a positive relationship with quality (higher costs mean higher quality). On the other hand, reactive costs arise from the use of resources reactively after a given quality has been produced or delivered. Reactive costs are incurred to correct poor quality and/or to compensate the affected parties (e.g. consumers). Therefore, reactive costs hold a negative relationship with quality (higher quality means lower cost). The bottom line of the framework is the following: a decision-maker can spend money to improve quality and, the higher the quality, the lower the cost of the consequences associated with it.

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. In other words, depending on the context of study, the cost-quality relationship may or may not be contemporaneous. 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). Similarly, 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 spillovers beyond period *t*. The framework recognises the generic linkages between quality and costs, and the potential for multiple temporal dimensions in which these linkages can exist (*t, t+1,… t+n*). The temporal dimension will highly depend on the context of study.

Defining proactive and reactive resources

The underlying assumption in all models of cost of quality is that poor quality have consequences. This is precisely what generates a two-way relationship instead of a simple positive relationship where more costs mean more quality. We argue that it is important not to take this for granted and understand what lies behind. It is crucial to understand why there may be consequences and what the consequences are. First, there can be consequences for various reasons. For example, if the level of quality of a product or service is linked to explicit promises or implicit customer expectations (e.g. product warranty; a train ticket that promises an arrival at certain time), or if an asset quality deterioration makes its repair much more costly. Hence, consequences can be explained and can be related to different factors, e.g. to the expected level of quality or to the life cycle of the good.

Secondly, the application of the framework should define, for each case study, what the consequences are. In some cases there may be aspects of quality where poor quality does not have any costly consequences. In other cases, the consequences can go beyond easily observable costly consequences such as refunds to customers, and also include long-term demand reductions that are more difficult to observe. Similarly, if we accept that demand effects can be related to quality, a firm that deliberately sacrifices part of the production to ensure greater quality (i.e. quantity-quality trade-off) is also making a proactive investment in quality in a more subtle way by giving up some profits. The limits of the framework can be narrower or wider depending on the context of study.

Related to what the consequences are, it is also important to specify who the decision-maker is. The framework does not specify it. For example, we can take the perspective of a private firm or the perspective of a regulator interested in social welfare. The former will probably focus on the costs incurred by the firm, whereas the later will also include any additional costs incurred by other agents (e.g. customers). The framework is valid for any perspective, and it is only a matter of defining the decision-maker and consequently what counts into the proactive and reactive cost categories.

Formalization of marginal cost of quality

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:

$$MCq = MPCq + MRCq \qquad (1)$$

Where:

$$MPCq > 0$$

$$MRCq < 0$$

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.

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.

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

Since MRCq is always negative and hence becomes smaller as quality increases, the overall effect is the 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 would be associated with inefficient choices (i.e. departures from rational behaviour). Economically efficient decision-makers should indeed face a positive MCq.

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). Figure 4 provides a graphical representation of the framework in terms of marginal costs and total cost of quality:

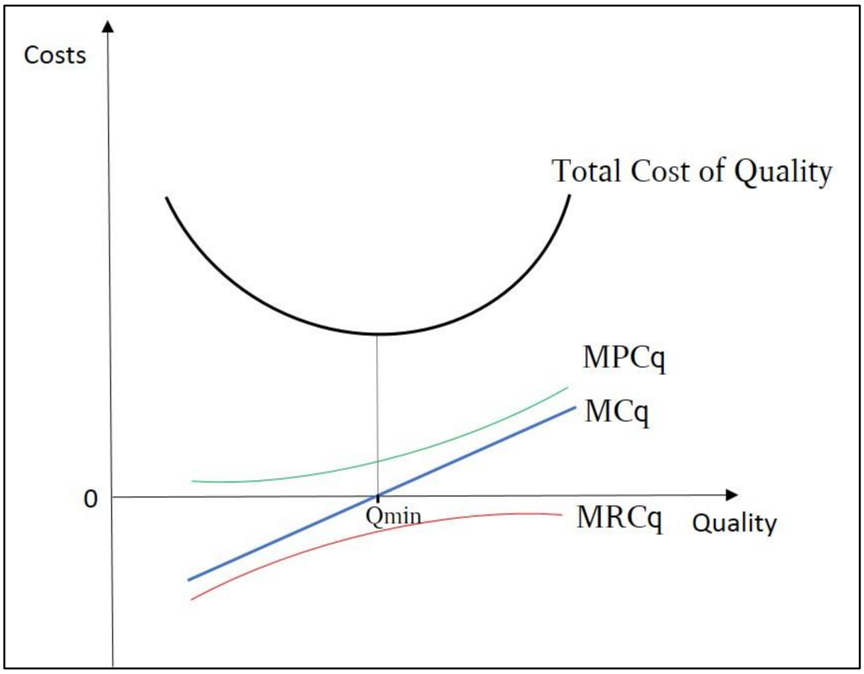


Figure 4 Costs and Quality theoretical framework, from a marginal cost perspective

The intuition of the framework is straightforward. 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 worsens also brings an increase in cost due to high reactive costs (MPCq < -MCRq, hence MCq<0). On the other hand, quality levels greater than *Qmin* would only be achieved at an extra cost: MCPq > -MCRq, hence MCq>0. The decision-maker should therefore consider the marginal benefit of the increase in quality to judge whether the positive associated MCq is worth the investment. While optimality analysis requires 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. The framework provides researchers with a theoretical guidance to approach the study of marginal costs of quality in any particular context.

3.2 Application of the framework to the case of railway reliability

The application of the framework to a particular cost-quality study requires: i) identification of the context of study and agent decision-maker, ii) a definition of quality, iii) understanding of which proactive and reactive costs may be associated with quality, and iv) understanding of the relevant temporal dimension for the cost-quality links. In this paper, the context of our empirical application is the railway operation in the UK and the decision-making units are the TOCs.

Defining quality: travel time reliability

This paper deals with 'reliability of travel times'. Reliability of travel times is one element of the quality of service of the railway network (Van Oort, 2016; Abate et al., 2013; Estache et al., 2007). Conceptually, it is a measure that indicates how faithful the actual service provided is, in terms of departure and arrival time, in comparison with the advertised schedule. Technically, 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 uses the generic term reliability[3].

However, in vertically separated industries, reliability can be associated to more than one decision-maker: e.g. infrastructure manager or multiple operating companies.

This is a problem that need to be overcome, since we can only infer the marginal cost of improving quality if the decision-making unit is responsible for it. Luckily, the railway industry in the UK allocates every minute of delay recorded across the IM and TOCs based on their responsibility. In 2016, the IM (Network Rail) was responsible for approximately 60% of delays (although note that this includes weather related incidents), with the TOCs being responsible for the remaining 40% of observed delay minutes. Delay minutes is a measure of (un)reliability, i.e. an inverse measure of quality. It is the only measure that can be directly attributed to the decision-making unit choices and hence costs. All other

[3] There are subtle differences between the different terms which are not considered in this paper as they would not change the theoretical relationship between cost and reliability.

measures of reliability or unreliability are not disaggregated by who is responsible. Therefore, the remaining of the application will focus on this measure of reliability: delay minutes caused by TOC on self.

Proactive and Reactive costs related to travel time reliability

Abate et al. (2013) refer to the link between proactive costs and reliability as the "effort effect". Examples of proactive costs where a TOC can intervene with the aim of reducing delay minutes are: inspection, maintenance, renewals (preventive investment to prevent failures), rolling stock (larger vehicles fleet in case a failure occurs) or staff (appraisal and management). All these investment options can be observed within TOC costs.

Reactive costs are present in the railway for one main reason: reliability embodies 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 TOC 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 ever larger when affected customers are entitled to compensations from significant delays and cancellations.

Additionally, there will be interactions between demand and reliability which will not be reflected in TOC costs. For example, high delays can mean foregone revenue, which is a reactive cost which could be saved by investing in 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 into account.

In summary, for each TOC, we expect to observe both proactive costs and reactive costs associated with delay minutes. All TOCs should naturally have the incentives to run services such that MC of reducing delays is positive (i.e. locate above Q_{min}, as depicted in figure 4). In other words, efficient companies should face a positive marginal cost of quality – under an effective industry structure. Additionally, incentives to invest in quality beyond this minimum point would depend on demand and the marginal benefits of reliability improvements. 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 increases in reliability, since increases in reliability do not necessarily have to come at an extra cost (if firms produce quality below *Qmin*). But, also, because there are natural gains in moving beyond *Qmin*, regulators should also address and remove any industry constraints that prevent TOCs from operating where quality > *Qmin*.

Temporal dimension of the cost-reliability relationship for TOCs

The conceptual framework from Figure 3 shows that cost-quality effects may or may not be contemporary. In the health sector, some quality-costs effects can happen almost immediately. If some patients are discharged soon, hospitals save (proactive) costs. However, in some cases, the sooner they get discharged, the higher the probabilities they can come back. If they do come back, the (proactive) costs initially saved may well be spent anyway (but now, as reactive costs and with associated lower quality). In the electricity sector the cost-quality links can take longer to occur. Jamasb et al. (2012) hypothesise and show empirically that proactive costs have a stronger effect the year after the investment. Their results also show a lack of contemporaneous effect.

In the context of railway operation, certainly some reactive costs will happen immediately if reliability is low: e.g. staff extra hours and penalties paid to customers. Quality effects will also be contemporaneous for some proactive costs such as hiring staff to perform more regular inspections and maintenance or to improve management. Hence, we hypothesise a strong contemporaneous link between costs and reliability for TOCs operations. However, we do not discard the possibility of spillovers across years: e.g. TOCs can invest today in future additional rolling stock capacity.

In contrast, the context of railway track provision by the IM can be different. When the operation is separated from the provision of tracks (i.e. vertically separated railway), the relationship between costs and quality for the IM can face more significant spillover effects over the years. This is because the maintenance and renewals of tracks are more complex than that of rolling stock. It is more difficult and expensive to make significant changes in the short term. The network conditions are highly dependent on previous years' conditions, and so is quality. Proactive investments (or the lack of them, leading to reactive costs) can influence quality within a longer time frame. Thus, we would expect the links between costs and quality not to be as contemporaneous as for the operating counterpart. The IM context remains outside the scope of this paper and is being addressed in a separate paper (Gillies-Smith et al., forthcoming).

4. Methodology and data

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).

The model assumes that all controllable costs (C_{it}) of decision-maker *i* in period *t* are explained as a function of outputs (y_{it}), a set of input prices (p_{it}) and the level of quality achieved (q_i). Other explanatory variables of the model are: a time trend (t) and a set of variables accounting for heterogeneity among TOCs (z_{it}).

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

In our application, C_{it} are the total costs of train operating companies excluding any transfers to the Infrastructure Manager (i.e. access charges and performance penalties/compensations). This means Schedule 8 compensations – incentive payments related to punctuality – are also excluded from the total costs. As mentioned before, these payments constitute an incentives system with the IM and can be positive or negative (i.e. compensation from IM's poor performance). These are different from TOCs compensations to passengers. In order to estimate the actual cost of improving quality, and be able to contrast this with existing S8 incentives, the incentives payments must be excluded from the modelling. 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 self, since these effectively represent quality under TOC control. Delay minutes is an inverse measure of quality: more delays equal lower quality. Under this specification, the elasticity of cost with respect to quality can be calculated as (Jamasb et al., 2012):

$$\varepsilon_q = -\frac{\partial C}{\partial q} \qquad (3)$$

The marginal cost of reducing minutes of delay caused on self for TOC *i* in period *t* can be calculated as:

$$MC_q = \varepsilon_q \frac{C_{it}}{q_{it}} \qquad (4)$$

Where ε_q is the elasticity of the own-caused minutes of delay, equal to $-\,\partial C_t/\partial q_t$ in our model, and q_{it} is the own-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 impossible to separate contemporaneous marginal proactive costs and marginal reactive costs of quality. Only the overall marginal cost of quality (MCq) -i.e. of saving 1 minute of delay- can be estimated. Theoretically, we assumed MCq to be a combination of 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 q_{it} (reducing delays is costly overall). However, $MC_q < 0$ would also be possible if firms were not efficient or if there were constraints in the industry structure.

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 spill over reactive effects. If any, 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 compensations to be 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 and marginal costs via interaction terms between TOC characteristics and q_{it}.

Finally, it is worth noting that a non-significant coefficient on q_{it} would not mean a lack of relationship between cost and quality for a given TOC. It can also be interpreted as zero MCq.

Dataset

The model is applied on a panel data from eighteen train operating companies in the UK over a 5-year period, from the financial year 2010-2011 to 2014-2015. 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. Additionally, route-km is also included in the estimation to distinguish between scale and density effects (Smith and Wheat, 2012). The set of input prices contains the average salary in the firm and an average price per unit of rolling stock. Other variables (z_{it}) included in the models are load factor per vehicle (passenger-km/vehicle-km), the type of service dummies (intercity, regional and LSE) and the delay minutes caused by TOC on self. The descriptive statistics are shown in table 1 below.

Table 1. Descriptive statistics

Variable	Description	Mean	Sd.Dev.	Min	Max
EXP	*Total controllable expenditure, i.e. excludes track-access charges; in million £ (£,2004)*	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*	29.9	14.8	6.1	63.8
PASSKM	*Passenger-km*	3375.9	2143.8	545.1	8691.0
ROUTEKM	*Route-km*	1355.1	854.3	115.5	3065.8
DENS	*Density (train-km/route-km)*	0.029	0.016	0.012	0.056
STATIONS	*Number of stations operated*	143.6	127.2	0.0	464
LOAD	*Average Passengers per train (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 (LOAD/AVLEN)*	22.2	4.0	16.4	33.2
DMT	*Delay minutes caused by own TOC*	199461	162820	14065	672567
INTERCIT	*Intercity category (% of intercity services of TOC). Omitted category: regional services*	0.4	0.4	0.0	1.0
LSE	*London and South East category (% of LSE services of TOC).*	0.3	0.5	0.0	1.0

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.

5. Results and discussion

This section outlines the econometric results and their application to calculate elasticities and marginal costs of quality. After an extensive specification search, our preferred cost model is a restricted version of the translog functional form. We retained only a subset of the squared and interaction terms from a full translog specification, which was not supported by the limited data. This is not uncommon with full translog models (e.g. Morrison, 1999). 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.

The model estimates are shown in table 2. All variables are specified in logs after being scaled by the sample geometric mean. 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).

The preferred model uses a fixed effects (FE) approach. FE seemed more appropriate than a Random Effects (RE) approach due to the need to isolate the *within-firm* temporal variation in cost in relation to quality (see e.g. Jamasb et al., 2012). The FE specification includes a constant for each firm, which enhances the ability of the model estimates to pick up *within-firm* effects as opposed to differences across firm. This FE choice was also supported empirically: the Hausman test rejected the RE specification in favour of FE.

Table 2. Preferred Model Results

	Model 1				**Model 2 (incl. NR delays)**			
Variable†	**Est.**	**Std.Error**		**t-rat**	**Est.**	**Std.E.**		**t-rat**
SAL	0.497	0.027	***	18.18	0.507	0.027	***	18.91
ROUTKM	0.578	0.178	***	3.24	0.511	0.180	***	2.84
DENS	0.774	0.168	***	4.62	0.872	0.160	***	5.46
STATIONS	0.311	0.309		1.01	0.359	0.319		1.13
LOADVE	0.210	0.119	*	1.77	0.282	0.113	**	2.5
AVLEN	0.358	0.147	**	2.43	0.469	0.144	***	3.27
T	-0.101	0.046	**	-2.19	-0.098	0.047	**	-2.11
T^2	0.004	0.002	**	2.26	0.003	0.002	**	2.11
LOADVE^2	0.698	0.191	***	3.66	0.764	0.177	***	4.32
DMT	-0.077	0.029	***	-2.69	-0.088	0.035	**	-2.5
DMT^2	0.036	0.013	***	2.66	0.017	0.012		1.37
DMT*STA	-0.030	0.012	**	-2.54	-0.032	0.011	***	-2.79
DMT*SAL	0.053	0.021	**	2.53	0.066	0.021	***	3.18
DMT*LEN	-0.135	0.051	**	-2.64	-0.218	0.052	***	-4.21
DMT*LVE	0.259	0.077	***	3.37				
DMT(NR)					0.002	0.028		0.08
DMT(NR)^2					0.021	0.010	**	2.15
DMT(NR)*LEN					0.104	0.064		1.64
DMT(NR)*LVE					0.287	0.069	***	4.19
DMT(NR)*DENS					0.021	0.041		0.51
R-squared	0.999				0.999			
Observations	85				85			

, **, * indicates significance at the 90%, 95% and 99% confidence level respectively.*

†*All variable names correspond to those described in Table 1, except for T which is the time trend.*

First, we discuss the overall performance of the model. Secondly, we will address the results on the quality elements. All output and input price coefficients are estimated with the expected sign. The majority of coefficients are statistically significant. A Wald test did not reject that the coefficient on Stations is jointly significant with others, even with the least significant coefficient LOADVE. The FE approach implies that we are capturing part of the variation across firms via TOC-specific dummies (not reported). Hence, to some extent, it can be argued that most of the variability observed is inter-temporal variation. Since the data only contains 5 years of information, we must be careful in comparing the results with other literature, especially when most related works have used a RE approach (e.g. Smith, 2006; Smith and Wheat, 2012). For example, similar models using RE have been used to investigate economies of scale and density. It is not possible to compare our estimates of economies of scale and density with those. However, the results are somewhat within the expected range.

The first-order coefficient on delay minutes (DMT) is negative and significant at the 99% level of confidence. This means that, at the sample average, the cost elasticity of delay minutes is negative, i.e. reducing delays is associated with higher costs. As noted in the methodology section, delay is an inverse measure of quality. 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. This means the cost elasticity of quality is positive at the sample average, which means it is costly to improve quality. In other words, the average firm faces a positive marginal cost of quality (MCq > 0).

Additionally, the model was not constrained to estimate a constant elasticity of quality. The results show the elasticity varies in a number of ways for a given TOC and across TOCs. First, the squared term takes the reverse sign than the first-order coefficient, implying a quadratic relationship. This result matches the theoretical shape of the cost-quality relationship. Secondly, there is plenty of 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. This heterogeneity implies that each TOC faces a slightly different cost-quality relationship. More precisely, at the sample mean:

i) MCq is higher if trains are longer. This result may reflect that proactive costs such as preventive inspection and maintenance are higher if a train is made up of more vehicles.

ii) MCq is lower if salaries are higher. Staff costs can be both proactive and reactive (e.g. staff extra hours). This result shows that the reactive component has relatively more weight, hence improving quality is cheaper when salaries are high because a TOC will be saving delay related staff costs.

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 savings in passengers' compensations.

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. It is not clear how the number of stations managed can affect the cost-quality relationship. However, further modelling tests[4] show that this relationship might be driven by one TOC only which is somewhat an outlier in this respect as it does not control any stations at all. For some reason, this particular TOC faces a negative MCq, but it remains unclear whether this is due to not being in charge of any stations. This TOC, even though they could save costs

[4] Additional models not reported are available from the authors on request.

by reducing delays, there seem to be constraints that prevent them from doing so. One hypothesis may be that such constraint is precisely the lack of stations management, which might limit the management possibilities of the TOC. But this is highly hypothetical and we cannot test this as there is no other TOC with zero or even just a few stations managed.

Cost elasticity with respect to quality

We calculate the cost elasticity of quality (ε_q) for each TOC and plot it against different scales of quality: a) delay minutes (the raw variable used for estimation), b) delay minutes per train-km, and c) delay minutes per train planned. We do this since the absolute measure of delay minutes is not necessarily the same quality scale for every TOC. For example, 200,000 minutes of delay per year (the sample mean) are not the same for a TOC which runs only 6 million train-km than for another than runs 60 million train-km: for the former, this figure is likely to reflect bad quality, but it may reflect good quality for the latter. The graphs are shown in figure 5. Note that the horizontal axis (delay minutes) has been reversed in order to reflect quality. This allows us to compare the empirical results with the theoretical framework.

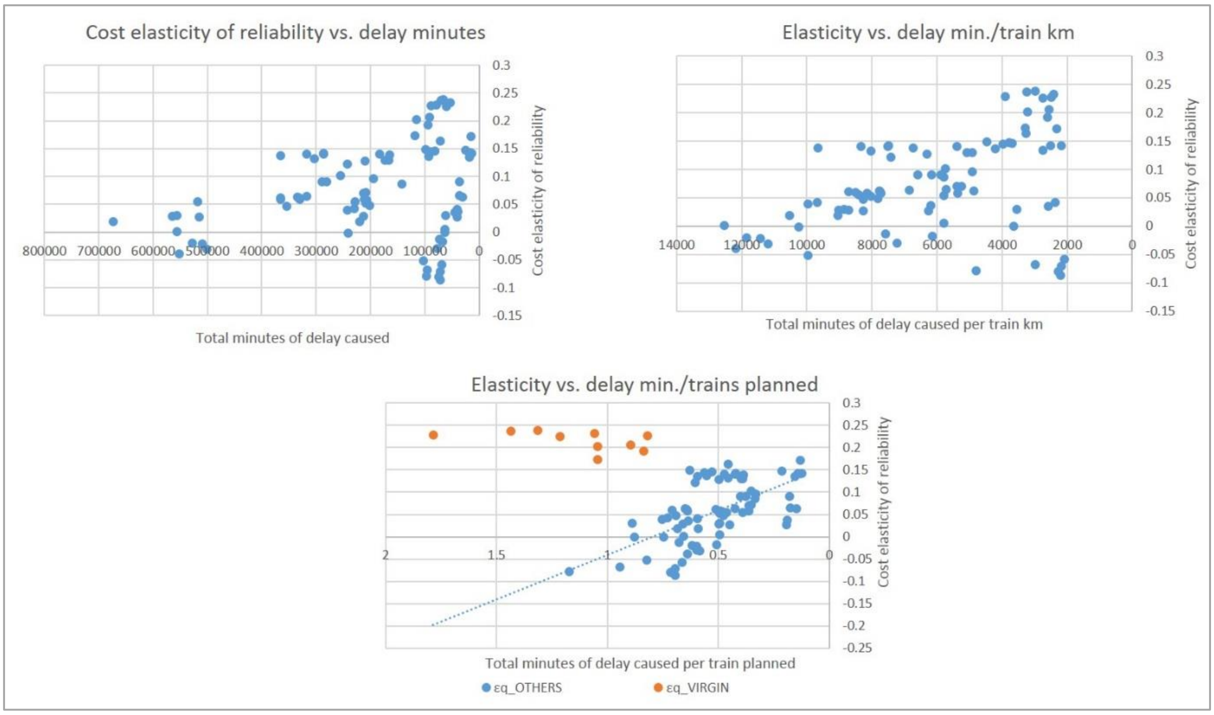


Figure 5. Estimates of cost elasticity of quality

In general, the elasticity increases with quality, showing that increasing quality becomes more expensive when quality is already high. This pattern of the elasticity of quality is consistent regardless of the scale of quality used, and is in line with the theoretical framework. In two of the graphs, those negative values slightly outside the trend correspond to the TOC with no stations and may be regarded as outliers. If we define quality as delays/train planned, the estimated results may suggest a substantially different curve for two TOCs, which happen to be the fastest intercity services running on two core corridors of the network (East Coast and West Coast franchises).

The possibility of cost-quality spillovers over time was tested using leads and lags on the delay minutes variable, in line with the specification employed by Jamasb et al. (2012). None of the leads or lags were found to be significant, hence the cost and quality relationship is highly contemporary in the context of train operating companies as expected. In other words, both proactive and reactive costs of quality occur within the same time period *t*.

Marginal cost of quality estimates

The marginal cost of quality for each TOC *i* in period *t* can be obtained multiplying the calculated elasticities of quality by the average cost of quality. The average cost of quality is calculated by dividing the total costs of TOC i in period t by the number of delay minutes (as in equation 4). The estimates are reported in figure 6 using the same horizontal axes than for the elasticity graphs of figure 5.

Additionally, it is possible to plot these MCq jointly with estimates of the marginal benefit of improving quality; we do so for the last graph in figure 6. We calculate the marginal benefits as the willingness-to-pay (WTPq) for quality improvement, i.e. reducing delays for rail passengers, using the results from the most recent and largest study on time and reliability valuation in the UK (ARUP et al., 2015)[5]. Since these WTPq per delay minute are available per passenger, and 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*.

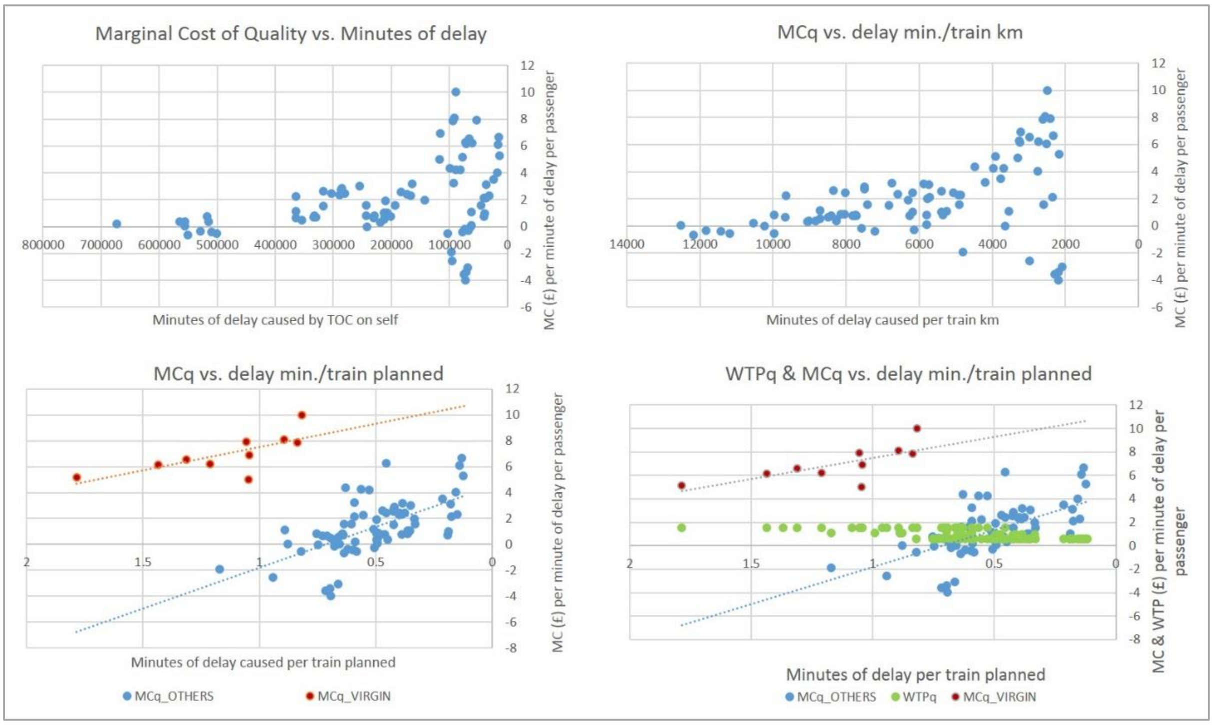


Figure 6. Estimates of Marginal Costs of quality (MCq)

[5] We use the behavioural values of travel time saving for rail travellers and the lateness multipliers to convert standard minutes of time into delay minutes of time. This is the standard approach in the literature (see e.g. Wardman and Batley, 2014). To cope with variation by purpose in the values, we apply weights based on the observed purpose split in the railway (with one exception: for Intercity TOCs, we reduce the commute share and extend the leisure share).

The graphs and interpretation of the distribution of MCq is analogous to that of the elasticities, reported previously. It should also be noted that there is an added layer of uncertainty in that MCq combines two estimates: elasticities and average costs. In a multi-output context, the calculation of the average cost is less intuitive than if only one output was produced. Nevertheless, when a holistic approach is employed in the cost model (including all inputs, outputs and costs) then it is still possible to compute average costs and marginal costs as indicated in equation 4 (see e.g., Darrough and Heineke, 1978).

The MCq is, on average, just below 2 £/min per passenger. The distribution of MCq ranges from -4 £/min to 10 £/min, although the extremes may be seen as outlier values. The vast majority of estimates range from -0.5 to 6 £/min. It is difficult to discuss the precision of these estimates for various reasons. The MCq is a subtle concept, possibly difficult to grasp even for TOCs themselves. Econometric techniques provide a way to estimate it, but it is difficult to relate the numbers to TOC activities and quality-related decisions. This is the case for output marginal cost estimation, but even more so for quality marginal cost estimation because quality is intertwined with multiple firms' decisions. Also, there is no previous evidence of MCq in the railway and therefore we cannot compare our figures with other sources. Consequently, we urge these estimates to be interpreted very carefully and highly encourage more work on this area.

After acknowledging the limitations of the approach, we now turn into discussing the MCq estimates. Regardless of the precise magnitude of these, the MCq sign is driven by the estimated elasticities. Therefore, the empirical MCq resemble the theoretical expectations and mostly fall within the right-hand side of the graph in figure 4 (MCq >0). This means that most companies are, at least, producing a quality level greater that Qmin, as expected. However, in roughly 18% of the cases[6], quality is lower than Qmin meaning that TOCs could improve quality and reduce costs. In these cases, TOCs are making quality decisions somewhat inefficiently.

There are several possible explanations 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 reactive cost savings. Hence, one naive explanation is that a TOC in this position has not realised this possibility. Understanding and achieving an optimal balance between proactive and reactive costs might not be an easy task. 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 and only after failing they find themselves expending more reactively (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, although a TOC realises the potential for cost savings and quality improvement, it is limited in its actions or budgets and cannot possibly undertake the small investment that would unlock greater reactive cost savings. It is outside of the scope of this paper to investigate what potential barriers would look like, but some candidates might be: congested network, tight timetables and any other managerial limitations (e.g. lack of stations control).

[6] Note, however, that for some of the small share of negative estimates, the negative elasticities are not significantly different from zero and hence these TOCs may technically be just around Qmin.

Finally, contrasting the MCq with the WTPq is useful to give some context to our MCq estimates. The comparison shows that the estimated average and range of MCq is not far from the WTPq. On average, the MCq is slightly higher than the WTPq. This indicates that in most cases it is more expensive to reduce delays than what people have declared to value the reduction. Note that the WTPq have been derived from asking people how much they would be willing to pay to receive more punctual services. However, it can easily be argued that these WTPq represent an underestimation of the real value of delays, since passengers could have felt that it is not their responsibility to pay for reliable services. In other words, we do not know if respondents felt that what they already pay for their tickets entitled them to reliable services.

Furthermore, one must be careful to perform welfare analysis based simply on the reported figures. These cost and benefit estimates do not represent the full picture and we argue they should not be used (if isolated) for social welfare analysis: the MCq of improving a delay is calculated at the individual firm perspective, but omits additional costs such as delay propagation to other TOCs, freight companies or to the IM. Also, for a social welfare analysis, additional long term effects on demand should also be accounted for and is far from clear than these are represented in the WTPq. The use of the MCq provided can be used as an input for welfare analysis only as part of a more comprehensive analysis.

Implications for performance incentives

One of the aims of the paper was to provide valuable information for the design of performance incentives in vertically separated industries. Both the theoretical framework and the empirical results may be used as inputs by rail regulators and IM alike. In this section, we briefly discuss the implications of the paper for quality incentives. However, a comprehensive discussion of incentives would need to consider all agents and not only TOCs perspective. This paper is an input for such a holistic approach which is not the scope of this work.

The main implication of the paper emanates from the theoretical work. 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 certain level of punctuality, e.g. via monetary compensations between TOCs and IM. 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 figure 4), as otherwise it would face higher (reactive) costs.

The design of performance incentives should consider natural incentives and the level of *Qmin.* Herein, our results show that it is possible to estimate Qmin. In the current circumstances, Qmin is very different across TOCs. Incentives might only be needed above and beyond *Qmin,* as otherwise there is a risk of subsidising an inefficient firm (e.g. Gutacker et al., 2013). On the other hand, the data we observe reflects TOCs choices under the current performance incentives regime and we cannot know what quality levels they would be achieving naturally (without the imposed incentives).

Another implication is the following: reactive costs, and hence the natural incentives, depend on the consequences faced by TOCs. One interesting aspect of this is the compensations to passengers for the delays suffered. These compensations are somewhat a hybrid between natural and imposed incentives. In the UK, these are not part of the official incentives regime (Schedule 8), hence we may argue they are considered as natural. However, the extent to which companies are obliged to compensate passengers vary and is regulated. In the UK, there is a threshold of delay minutes (30 minutes)[7] 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 achieve higher levels of quality simply by ensuring these reactive costs are incurred. 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 figure below, movement from Qmin_0 to Qmin_1).

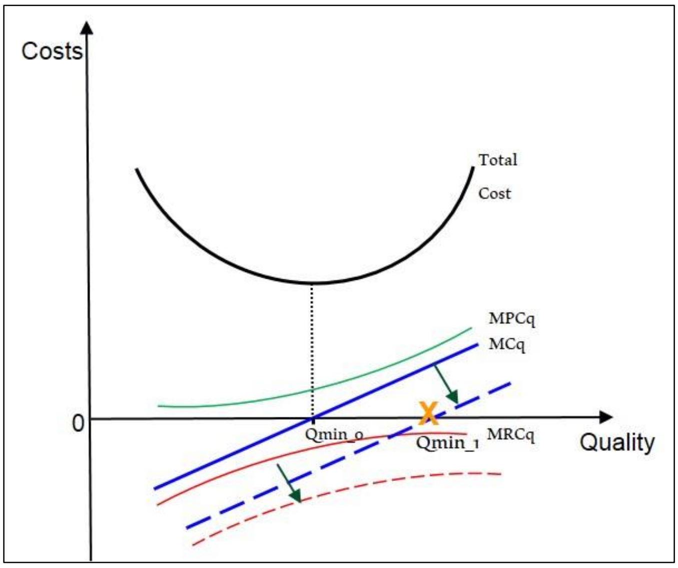


Figure 7. Effect of reinforcing delay consequences

The graph depicts the effect of facilitating compensation payments to passengers. Since this increases the reactive costs of TOCs, when they contemplate a quality improvement they would also observe a higher reactive cost saving (the MRCq curve shifts downwards). Hence, TOCs would therefore observe, for a given level of quality, a lower MCq for further improvements (MCq curve shifts downwards as MCq = MPCq + MRCq). With the new reactive costs curve after facilitating compensations, TOCs would have to consider the higher reactive costs.

However, in practice, this would not be so straightforward because TOCs are not responsible for many of the delays, but yet they have to compensate passengers for all. Regulators might consider keeping passengers' compensations out of the imposed incentives, but also making TOCs responsible only for those which they cause themselves (i.e. making sure the IM or other TOCs refund those external delay compensations to the paying TOC).

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

6. Conclusions

This paper has explored the relationship between travel time reliability and costs of train operating companies. In this process, our work makes a number of additional contributions. First, the paper has brought together several bodies of literature which had remained disconnected from each other. We have uncovered a central theme 'costs and quality' that applies across literature 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; management research has widely covered the topic developing Cost of Quality models from 1951, 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 too specific and remained unaware and disconnected from existing Cost of Quality models.

Secondly, we developed a generic 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). In line with the existing literature, the quality-cost relationship is marked by countervailing forces: proactive and reactive costs. Our framework proposed that marginal cost of quality is a combination of marginal proactive costs and marginal reactive costs. The framework was then applied to the context of train operating companies in the UK, but remains general enough to be used in other transport and non-transport contexts.

Third, we provided the first estimates on cost elasticity and marginal costs of reducing delays for train operating companies. The empirical work has shown that in most cases TOCs have been facing a positive marginal cost of quality. At the sample average, we found that it is costly to reduce delays for TOCs. This was expected given that TOCs operate under a performance incentives regime. The cost elasticity of quality was on average around 0.07, but it varied within and across TOCs. TOCS characteristics like average length of trains, vehicle load or salaries were found to influence the estimated elasticities. Also, the marginal cost of quality increased with the level of quality, as theoretically expected. This heterogeneity also revealed that the marginal cost of quality was also not significantly different from zero or negative in approximately 18% of the cases. In those cases, reducing delays would also bring cost savings, and it remains unclear why TOCs would not exploit such opportunity. One possible explanation might be the presence of industry structure constraints that prevent TOCs from doing so. Overall, the results were highly consistent with the theoretical framework.

This paper is only a first step to improve our understanding of quality in railway supply. Both the theoretical framework and the empirical application can be used to aid the design of incentives systems and industry structure in relation to quality aspects. Travel time reliability is one aspect of quality, but the framework can easily be translated to other quality contexts. Finally, new empirical evidence would be highly welcome to contrast and complement the first estimates of marginal cost of delay reductions provided in this paper.

7. References

Abate, M. Lijesen, M., Pels, E. and Roelevelt, A (2013). The impact of reliability on the productivity of railroad companies. Transportation Research 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', Journal of Business research, Vol. 67, pp. 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 05/05/16.*

Carrion, C. and Levinson, D. (2012). Value of travel time reliability: A review of current evidence. Transportation Research Part A-Policy and Practice. 46(4), pp.720-741.

Coelli, T.J., Gautier, A., Perelman, S. and Saplacan-Pop, R. (2013). Estimating the cost of improving quality in electricity distribution: A parametric distance function approach. Energy Policy. 53, pp.287-297.

Crosby, P.B. (1979). Quality is Free. New York: McGraw-Hill.

Darrough, Masako N. and Heineke, John M. (1978). Multi-Output Translog Production Cost Function: The Case of Law Enforcement Agencies. *In* Economics Models of Criminal Behaviour, John Heineke, ed., North-Holland Publishing Company.

Estache, A., Perelman, S., & Trujillo, L. (2007). Measuring quantity-quality trade-offs in regulation: the Brazilian freight railways case. Annals of public and cooperative economics, 78(1), 1-20.

Feigenbaum, A. V. (1956). Total quality-control. Harvard business review, 34(6), 93-101.

Freeman, J.M. (1995). Estimating quality costs. Journal of the Operational Research Society, 46, 675–686.

Kennedy, J and Smith, A.S.J. (2004). Assessing the Efficient Cost of Sustaining Britain's Rail Network. Perspectives based on Zonal Comparisons. *Journal of Transport Economics and Policy, Volume 38, Part 2, pp. 157-190*

Gillies-Smith et al., forthcoming, PhD Thesis, 2017.

Gutacker, N., Bojke, C., Daidone, S., Devlin, N., Parkin, D. and 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: 931–947*

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 6th July 2017

Hwang, G. and Aspinwall, E. (1996). Quality cost models and their application: a review. Total Quality Management, 7(3): 267-282.

ITS and MVA (2012). Demand impacts of recovery time on railway timetables

Jamasb, T., L. Orea and M. Pollitt (2012). "Estimating the marginal cost of quality improvements: The case of the UK electricity distribution companies." Energy Economics 34(5): 1498-1506.

Juran, J. M. (1951) Juran's Quality Handbook, 1st edn (New York: McGraw-Hill).

Juran, J.M. and Gryna, F.M. (1988). Quality Control Handbook, 4th Edn, New York: McGraw-Hill

Juran, J.M., Seder, L.A. and Gryna, F.M. (1962) Quality control handbook, 2nd ed. McGraw-Hill Book Company, New York, NY

Link, H. (2017). An efficiency analysis of rail passenger subsidies in Germany combining perceived and objective quality of service indicators and conventional output. Paper presented at ITEA 2017, Barcelona.

Masser, W.J. (1957) The quality manager and quality costs, Industrial Quality Control, October, pp. 5-8.

Merkert, R., Assaf, A.G. (2015): 'Using DEA models to jointly estimate service quality perception and profitability – Evidence from international airports', Transportation Research Part A, Vol. 75, pp. 42-50.

Morrison, C.J. (1999). Cost structure and the measurement of economic performance: productivity, utilization, cost economies, and related performance indicators. Boston, MA ; London: Kluwer Academic Publishers.

Mutter, R. L., Greene, W. H., Spector, W., Rosko, M. D., & Mukamel, D. B. (2013). Investigating the impact of endogeneity on inefficiency estimates in the application of stochastic frontier analysis to nursing homes. Journal of Productivity Analysis, 1-10.

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

Office of Rail Regulation (2013). Rail passengers compensations 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 6th July 2017

Peimbert-Garcia, R.E., Limon-Robles, J. and Beruvides, M.G. (2016). Cost of quality modelling for maintenance employing opportunity and infant mortality costs: An analysis of an electric utility. The Engineering Economist. 61(2), pp.112-127.

Peterson, E.B., Neels, K., Barczi, N. and Graham, T. (2013). The Economic Cost of Airline Flight Delay. Journal of Transport Economics and Policy. 47, pp.107-121.

Schiffauerova, A. and Thomson, V. (2006). A review of research on cost of quality models and best practices. International Journal of Quality and Reliability Management, 23(6): 647-669.

Smith, A.S.J. (2006). Are Britain's railways costing too much? Perspectives based on TFP comparisons with British Rail 1963-2002. Journal of Transport Economics and Policy. 40, pp.1-44.

Smith, A.S.J. and Wheat, P. (2012). Evaluating Alternative Policy Responses to Franchise Failure Evidence from the Passenger Ran Sector in Britain. Journal of Transport Economics and Policy. 46, pp.25-49.

Srivastava, S. K. (2008). Towards estimating Cost of Quality in supply chains. Total Quality Management & Business Excellence, 19, 193-208.

Van Oort (2016). Incorporating enhanced service reliability of public transport in cost-benefit analyses. Public Transp, 8, 143-160.

Wardman, M. and 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, 41:1041–1069

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 Journal, 1-48.

Funded under the H2020 programme

Collaborative project H2020-MG-2015-2015 GA-636237

Needs Tailored Interoperable Railway – NeTIRail-INFRA

Deliverable D1.7
Incentives Final Report Annex 3
Methodological Aspects of Marginal Cost Modelling: Estimating the marginal cost of different vehicle types on rail infrastructure

Document ID: NeTIRail-WP1-D1.7v1.0-FINAL – ANNEX3

Due date of Deliverable: 30/09/2017

Actual submission date: 21/12/2017

Dissemination Level		
PU	Public	x
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Task leader for this deliverable: Professor Andrew Smith, Institute for Transport Studies, University of Leeds

Document status		
Revision	Date	Description
V0.1		First draft (authors Andrew Smith, Kristofer Odolinski, Saeed Hossein Nia, Per-Anders Jönsson, Sebastian Stichel, Simon Iwnicki and Phillip Wheat)
V0.2	14/12/2017	Review by USFD and ALU-FR
V1.0	21/12/2017	Final version
Reviewed	YES	

Executive Summary

A combination of engineering and economic methods is used to estimate the relative cost of damage mechanisms on the Swedish rail infrastructure and marginal costs of different vehicle types. The former method is good at predicting damage from traffic, while the latter is suitable for establishing a relationship between damage and cost. The best features of both methods are used in a two-stage approach, demonstrating its applicability for rail infrastructure charging.

The estimations are based on 143 track sections comprising about 11 000 km of tracks. In the first stage, simulations based on engineering models are performed to predict the damage caused by different vehicles during 2014. Inputs in the simulations are ideal track geometry and track irregularities, vehicle speeds, wheel and rail profiles, and axle loads. Moreover, vehicle models in the simulations are chosen depending on the traffic that ran on the 143 track sections during 2014. The damage outputs from the simulations are measures on track settlement, wear of rail, rolling contact fatigue (RCF) and track component fatigue. The damage measures are used in the second stage of the approach, in which a statistical model is specified where maintenance cost is a function of the different damage measures as well as other cost drivers. The statistical model is estimated using information on actual costs during 2014, which results in cost elasticities with respect to the damage mechanisms. These elasticities indicate the relative costs of the damages. However, a strong correlation between the damages at the track section level makes it difficult to isolate the cost impact of each damage type. Still, the preferred model provides significant cost elasticities for wear of rails and track settlement. These damage types capture the cost impact from RCF and track component fatigue given the strong correlations between the damages on different track sections.

The cost elasticities for the damage mechanisms are used to derive the marginal cost per damage unit. Together with information on the amount of damage per ton-km each vehicle type has caused, a marginal cost per ton-km and vehicle type is calculated. The results show a substantial variation in the marginal cost per ton-km for different vehicle types running on the Swedish railway. However, this variation is mainly driven by the cost elasticities for wear and track settlement, as well as the differences in wear per ton-km and track settlement per ton-km between the vehicle types. Hence, the marginal costs per vehicle type do not reflect the vehicles' relative differences in RCF per ton-km and track component fatigue per ton-km.

All in all, this study demonstrates how the estimated relative costs of damage mechanisms can be used to calculate the marginal wear and tear cost of different vehicle types. The results are relevant for infrastructure managers in Europe who desire to differentiate their track access charges such that each vehicle pays its short run-marginal wear and tear cost, which can create a more efficient use of the rail infrastructure. More observations over time can be useful in future research in order to provide more reliable and robust estimates, as well as for isolating the cost impact from each damage mechanism.

Table of contents

Abbreviations and acronyms

Abbreviation / Acronym	Description
MC	Marginal cost
RCF	Rolling contact fatigue
ORR	Office of Rail and Road

1. Introduction

Operating a train service generates costs for the management of the rail infrastructure. Research on these costs became relevant for European policy after the vertical separation between infrastructure management and train operations in the 1990s, requiring track access charges to be set. To create an efficient use of the infrastructure, each vehicle should at least pay its short-run marginal cost, which is a requirement supported by EU legislation (see European Commission Directive 2012/34/EC).

One component of the costs incurred by a train service is the wear and tear of the rail infrastructure. The vertical force on the tracks created by the weight of the train is a crucial factor for this damage, and ton-km has been the most widely applied charging unit in Europe. However, the damage per ton-km can vary depending on the vehicle type used, where the number of axles and bogie type are important characteristics in this respect. Differentiating the track access charge with respect to variations in damage and cost incurred by different vehicle types creates stronger incentives for running more "track friendly" vehicles, and would create an even more efficient use of the infrastructure compared to a ton-km charge. Britain and Switzerland are examples of European countries that have chosen to differentiate their track access charges by vehicle type and ton-km. This type of charge requires an estimation of the marginal cost of different vehicle types running on the rail infrastructure, which is the purpose of this paper.

Different approaches have been used in the literature to determine the marginal wear and tear cost. The top-down approach tries to establish a direct relationship between costs and traffic using econometric techniques (see for example Link et al. 2008 and Wheat et al. 2009), while the bottom-up approach uses engineering models to estimate the damage caused by traffic. The damages are then linked to maintenance and renewal activities and their respective costs (see for example Booz Allen Hamilton 2005 and Öberg et al. 2007). A combination of these approaches has been proposed by Smith et al. (2017): a two-stage approach in which simulation methods (engineering models) are used in the first stage to estimate the track damage caused by the rail vehicles running on the tracks. The relationship between damage and costs are then established using econometric methods in the second stage.

The reason for combining the econometric and engineering approaches in this type of exercise is that they can complement each other. The strength of the former approach is that it uses actual costs and can put few restrictions on the elasticities of production. However, it has difficulties in picking up the complexity of the relationship between different vehicle types and costs – that is, it struggles to provide estimates by vehicles. The engineering approach is on the other hand good at predicting the relative damage caused by different vehicles, but has difficulties in linking the damages (caused by traffic) to actual costs.

Smith et al. (2017) applied their approach on Swedish data comprising 45 track sections, to demonstrate the feasibility of the method. In this paper, we apply the same two-stage approach with an aim to increase the precision of the marginal cost estimates. The contribution of this paper is therefore to test if the two-stage approach is applicable for charging purposes; can it be a viable approach for infrastructure managers that wish to differentiate their track access charges by vehicle type? To do so, we use a significantly larger dataset comprising 143 track sections in Sweden. Moreover, the simulation stage of the approach is refined, with significantly more detailed vehicle models in the simulation stage compared to the previous study. Hence, we make fewer assumptions on the damage caused by certain vehicles on the track. An extra damage mechanism is also included in our study.

The outline of the paper is as follows. The methodology is described in section 2. Sections 3 and 4 present the estimation stages in our approach in more detail. A description of the data is given in section 5. The estimation results are presented in section 6 together with a demonstration of the marginal cost calculations. Section 7 concludes.

2. Methodology

The econometric (top-down) approach and the engineering (bottom-up) approach are the two main methods to determine how wear and tear costs of the rail infrastructure vary with traffic. The former has become the most widely used approach, and its results are applied in many European countries. Munduch et al. (2002) and Johansson and Nilsson (2004) are early examples of studies that use a (translog) cost function to derive a cost elasticity with respect to traffic, using econometric techniques. This cost elasticity shows how a proportionate increase in traffic affects costs proportionately (the elasticity is multiplied with the average cost in order to get a marginal cost estimate). A set of control variables are included to account for heterogeneity in the production environment - that is, to isolate the effect traffic has on costs. This method has however not been successful in isolating the effect different vehicle types has on wear and tear costs, and the traffic measure used in econometric studies is generally gross tons.

Instead of trying to establish a direct relationship between traffic and costs, the bottom-up approach uses engineering models to determine the damage caused by traffic. The damage is often categorized as rolling contact fatigue, track settlement, or wear of the rail. The method can provide estimates of the damages caused by different vehicles, with the possibility to account for current infrastructure characteristics such as track geometry and rail profiles. These damages are then linked to costs to produce a marginal cost. This can be done with information on the volume of activities made to rectify the damages and the unit cost of those activities. Costs can then be allocated to different vehicle types. This approach has been used in Britain (see for example Booz Allen Hamilton 2005 and ORR 2013). However, a critical point in the approach concerns the relative cost of the different types of damages. For example, Öberg et al. (2007) use an engineering approach to produce estimates of the amount and type of damage caused by different vehicles on the Swedish railway network. The damage types are then assigned shares of maintenance and renewal costs based on advice from experts within the Swedish Rail Administration[1], which in turn enables a calculation of the marginal cost per vehicle type.[2] The experts' advice may, or may not be, close to the actual cost shares. In general, the link from damages to costs needs to account for external factors such as a heterogeneous production environment, which can vary in aspects that are difficult to capture without a statistical (econometric) approach. Examples are rail age (proxy for accumulated use) and track quality.

[1] This organization merged with the Swedish Road Administration in 2010, forming the Swedish Transport Administration.

[2] More specifically, based on the damages' cost shares, a calibration of cost coefficients of the damage types is made so that their model reflects an average marginal cost per ton-km on the railway network that was estimated in Andersson (2007) (see p. 56, Öberg et al. 2007). The cost coefficients are then used together with the simulated damages of the different vehicles to produce a marginal cost per vehicle type.

The approach proposed by Smith et al. (2017) is to use an econometric model to estimate the share of costs that can be attributed to the different damages - that is, the relative cost of the damage types. The same approach is used in this paper. More specifically, the estimation approach (depicted in Figure 1 below) consists of two stages. Similar to the bottom-up approach, we perform simulations based on engineering models in the first stage. We use traffic data together with infrastructure characteristics to simulate four different damage mechanisms: track settlement, wear of rails, rolling contact fatigue (RCF), and track component fatigue. Hence, we include an extra damage mechanism (track component fatigue) compared to the study by Smith et al. (2017). This damage mechanism may eventually require replacement of components and can be important to consider given that minor replacements are defined as maintenance.

The output from the first stage is measures of the different damage types per ton-km for each vehicle type. Apart from differences in traffic between track sections on the rail network, these damage measures can also vary for each section due to the distinct characteristics of the sections such as track geometry and curvature. The measures are then scaled up based on the traffic volume of each vehicle type on the different track sections. In that way, we produce measures on the total track component fatigue, track settlement, RCF and wear of rails, that traffic has caused on a section. We use these damage measures in the second stage, in which a statistical model is formulated where maintenance cost is a function of the damage mechanisms and other cost drivers. Cost elasticities are derived from the statistical model, giving us the relative cost of the damage types. Based on the information from the simulation, we can estimate the marginal cost of the vehicle types.

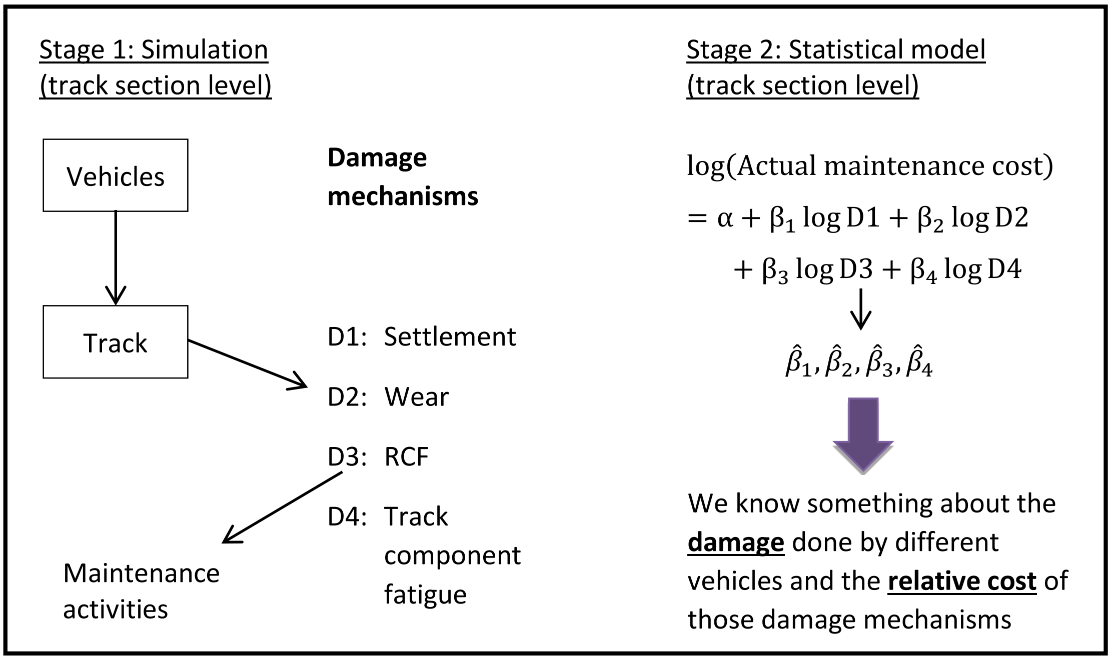

Stage 1: Simulation
(track section level)
Stage 2: Statistical model
(track section level)
Vehicles
Damage mechanisms
Track
D1: Settlement
D2: Wear
D3: RCF
D4: Track component fatigue
Maintenance activities
$\log(\text{Actual maintenance cost}) = \alpha + \beta_1 \log D1 + \beta_2 \log D2 + \beta_3 \log D3 + \beta_4 \log D4$
$\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4$
We know something about the damage done by different vehicles and the relative cost of those damage mechanisms

Figure 1: Overview of the methodology (revised figure from Smith et al. 2017)

A detailed description of the simulations and the econometric model we estimated is provided in sections 3 and 4 respectively.

3. First stage: simulations

Calculating the amount of track damage is a complex matter. First, it is a function of track quality itself - that is, whether

- the track is newly built or recently repaired,
- the sleepers are wooden or concrete, or if it is a slab track,
- the environment is humid or dry,
- wheel and rail profiles are well matched or not.

We account for the track quality in the simulations by using measurements on the track geometry, which will differ depending on the age of the track or if it has recently been repaired. However, we only consider concrete sleeper tracks and a constant environment (which creates a constant friction level) in the simulations. Moreover, we do not account for the actual matching of the wheel and rail profiles as it is beyond the scope of this paper. This means, only new unworn wheel and rail profiles are used.

The vehicles operating on the track are also of significant importance for the damage incurred. Some of these determining factors - that are considered in our simulations - are

- bogie design,
- wagon structure,
- axle load and
- vehicle speed.

The track damages investigated in this study are categorised into four different types. These are:

- track component fatigue,
- track settlement,
- rolling contact fatigue (RCF) and
- wear of rails.

We quantify and calculate the amount of track damage for each of the damage types listed above, using dynamic simulation and the damage prediction models available in the literature. The simulations are performed on 143 track sections in Sweden, which in total comprise about 11 000 km of tracks. Traffic data from 2014 are used to identify the vehicle types running on each track section. This includes information on the number of vehicles operating on each track section, as well as the vehicle types and their ton-km values. The dynamic simulations and the modelling issues are described in more detail in the following section.

3.1 Dynamic simulations

Computer based vehicle dynamics calculations using multibody simulation software have been widely used by companies and researchers for many years. These simulations are mainly used to predict the dynamic behaviour of the vehicles for different track conditions and operating conditions, thus making sure that the requirements on derailment, ride quality, track forces, RCF and wear will be met. It is also a very powerful tool to reduce the number of expensive field measurements. The modelling issues in these dynamic simulations are divided into three parts: track models, vehicle models and the wheel-rail contact.

We use a track model representing concrete sleeper tracks in the simulations, which is the sleeper type used on most of the tracks in Sweden (see Chaar and Berg 2006 for more information on track flexibility characteristics and its validation). The model of the track comprises of ground, ballast, rails and stiffness between these bodies as shown in Figure 2.

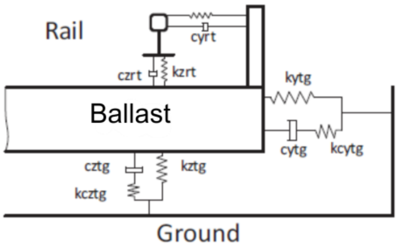


Figure 2: Example of model for track flexibility

In principle, vehicle modelling starts with defining the rigid or flexible bodies connected by springs, dampers and links. The bodies include car body (passenger wagons or freight baskets), bogie parts (frames, bolster beams, and possible steering links) and axles (axle box and wheels). A four-axle bogie vehicle may be modelled as seven rigid bodies which are shown in Figure 3. These rigid bodies may have all six degrees of freedom unless they are constrained. These six motions are vertical, lateral, longitudinal, pitch, yaw and roll.

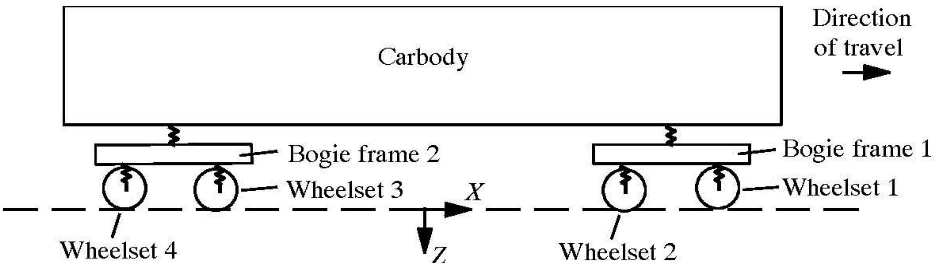


Figure 3: schematic model of a four axle bogie vehicle

For each rigid body mass, we have to know the moments of inertia, nominal positions of the centre of gravities, and locations of the coupling elements. The other important part is modelling the suspension elements. These elements are mainly springs, dampers and frictional contacts. Some of these elements are heavily non-linear and their behaviour depends on the applied loads, forces and displacements. A suspension element could be a coil or rubber spring, air spring, leaf spring, hydraulic damper, and metal wedge. For each type of element mentioned, there is a mathematical model which represents its dynamic behaviour.

The vehicle models we choose depend on the traffic running on the 143 track sections in this study. According to the traffic data, there were 111 rail vehicles in total operating on these sections in 2014. It is not possible to model each of these vehicles separately. Thus, the vehicles are categorized based on the type of the running gear, vehicle category (freight/passenger), axle load and maximum speed. The chosen categories are presented in

Table 1. Moreover, due to time restrictions, we only run simulations for vehicles that comprise more than 9 per cent of a track section's total ton-km. The vehicles that are left out are assigned the damage values from simulated vehicles with the most similar characteristics with respect to damage.

Table 1: Vehicle model categories with their maximum speed

Categories	Max. speed km/h
Motor coach 4x16 t*	200
Passenger car 4x14 t	160
Motor coach 4x16 t**	200
Motor coach 4x12 t*	140
Motor coach 4x21 t, high centre of gravity**	200
Motor coach with Jacob bogie 3x16.5 t**	160
Motor coach with Jacob bogie 3x12.5 t*	200
Freight loco 6x20 t	120
Freight loco 4x20 t	120
Freight loco 6x30 t	70
Passenger loco 4x19 t	140
Passenger loco 4x19 t	175
Freight wagon (2x22 t or 2x6.5 t)	100
Three-piece bogie 4x30 t	60 (laden)
Three-piece bogie 4x6.5 t	60 (tare)
Y25 bogie 4x22 t	100

* Flexible wheelset guidance, ** Stiff wheelset guidance

All the mentioned models are carefully designed and the results of the calculations are validated against the field measurements for certain types of the vehicles. To design and run the simulation models the Swedish multibody simulation software GENSYS (2015) is used.

Lastly, we need to model the wheel and rail contact, which is the tiny contact area between the wheels and the rails that is subjected to very high stresses. The way to calculate these stresses is crucial for prediction of the dynamic interaction between the vehicle and the track. A wheel-rail contact model consists principally of a wheel-rail geometry module, a creep/spin calculation procedure and a creep force generator. The theories are described for example in Andersson et al. (2015). In this study, the Hertzian solution and Kalkers FASTSIM method is used for the normal and tangential contact problem respectively.

3.2 Simulation inputs

Inputs needed for the simulation are ideal track geometry and track irregularities, vehicle speeds, wheel and rail profiles, and axle loads. Data on track geometry has been provided by the Swedish Transport Administration (Trafikverket), and originates from track measurements in 2014. The geometry includes the longitudinal position on the line, track super-elevation, track lift and track curvature. The irregularities include lateral, vertical, cant and gauge irregularities.

We set the vehicle speed as a function of cant deficiency in a way that maximum allowed cant deficiency can be reached, where the maximum lateral acceleration will be limited according to Banverket (1996). There are three categories defined based on the vehicles running gears,

- category A; conventional vehicle with older running gear $a_{y,lim} = 0.65\ {}^{m}/_{s^2}$,
- category B; vehicles with improved running gear $a_{y,lim} = 0.98\ {}^{m}/_{s^2}$,
- category C; X2000 and other high speed trains $a_{y,lim} = 1.60\ {}^{m}/_{s^2}$.

The maximum vehicle speed is limited with the permissible speed on each line.

As wheel and rail material will be worn gradually, their profiles' shape also changes. Therefore, it is almost impossible to use all the actual wheel-rail profiles in operation. However, as the majority of rails in Sweden have UIC 60 and the wheels have S1002 profiles, we use these profiles in the simulations. The rail inclination in Sweden is 1:30.

Apart from the Iron-Ore locomotives and wagons, which run at 30t axle load, the rest of the freight wagons are simulated with 22.5t axle load when they are fully loaded. The weight of the empty freight wagons is calculated based on their basket, bogie and other parts' design. For passenger trains, there is no generally accepted standard for calculation of the passenger loads. However, the weight of a passenger including hand luggage can be estimated to 80 kg. According to European standard EN 12663 (CEN 2010), the number of passengers in a coach is equal to the number of seats, which is the standard we use.

3.3 Track damage

We calculate four types of track damages for each vehicle on each track section: track settlement, track component fatigue, wear of rails and rolling contact fatigue (RCF).

Track settlement has a major influence on maintenance cost and is usually caused by high wheel-rail forces from passing vehicles. This type of damage depends strongly on the amount of track irregularities. Thus, axle load, unsprung mass and speed, track construction, track quality and track condition are the most important factors determining the magnitude of the damage. To calculate the settlement damage, various empirical models have been used. However, in most of them, the vertical wheel-rail force raised to a power is used as a damage indicator. In the present study the adapted TUM (Technical University of Munich) settlement calculation model is used:

$$Settlement = A\ \cdot Q^{1.21} \log \mathrm{N} \tag{1}$$

where,

N = number of axles passes

Q = Vertical force at the wheelset

A = constant; (A=1 in the current work)

Internal fatigue damage due to repeated loading is a function of both vertical and lateral track forces. The components affected by the repeated loading are rails, rail pads, rail fasteners, and sleepers. The calculation method is developed by UIC/ORE (1987) based on extensive tests and it is complemented by Öberg et al. (2007) with a lateral force component – that is, the resulting force on either rail.

$$Track\ component\ fatigue = \sum_{i=1}^{n_v} \sqrt{{Q_{tot\_i}}^2 + {Y_{qst\_i}}^2}^{\,3} \qquad (2)$$

where,

n_v= number of axles

Q_{tot_i} = total vertical force including quasistatic and dynamic forces

Y_{qst_i} = quasistatic lateral force

Wear of rail and wheel is a function of material properties (steel grade), contact pressure (axle load, wheel-rail profile), sliding velocity (creepage and spin), weather condition (sun and rain) and lubrication (track side or vehicle based). In this study, the friction level is assumed to be 0.45 for all the simulations unless the locomotives are equipped with vehicle based lubrication systems - that is, the Iron-Ore loco. To predict the wear on rails, several prediction models are proposed in literature (see Enblom 2004). One of the most widely used and simple ways to predict the amount of wear is to calculate the dissipated energy in the wheel-rail contact patch. This is based on an assumption that there is a linear relationship between wear and energy dissipation. Energy dissipation per meter running distance can be calculated as:

$$\boldsymbol{\overline{E} = F_x \nu_x + F_y \nu_y + M\varphi} \qquad (3)$$

where,

$\boldsymbol{F_x}$ and $\boldsymbol{F_y}$ are longitudinal and lateral creep forces,

$\boldsymbol{\nu_x}$ and $\boldsymbol{\nu_y}$ are longitudinal and lateral creepages,

M is the moment and φ is the spin in the contact patch.

In the present study, it is assumed that if the wear values are below 160 J/m, then the wear regime is mild wear and the value of wear damage is neglected (Smith et al. 2017).

To calculate surface initiated cracks due to RCF, again the energy dissipation based theory is used (see Figure 4). Here, first the energy dissipation is calculated and then the RCF index is picked accordingly.

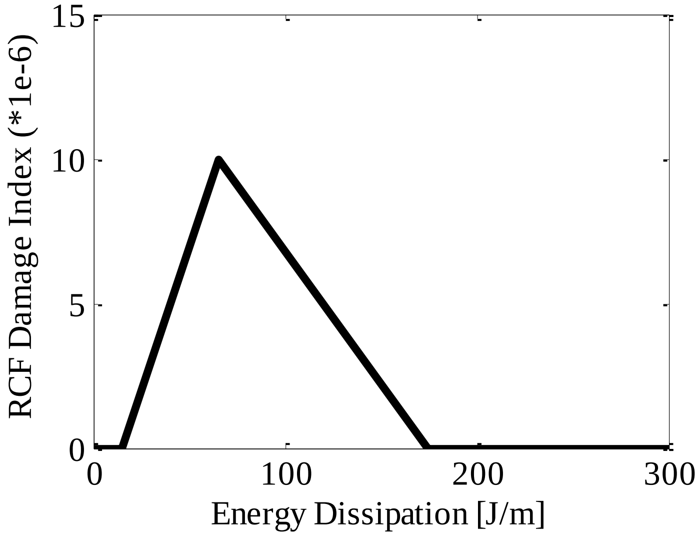


Figure 4: Rail RCF damage function (Burstow 2004)

3.4 Time domain analysis

All equations of motions must be integrated numerically in each time step. The results of each time step are the inputs for the next one. This is called initial value numerical calculations. In this study 1ms is used for the time steps. Depending on the track section length, the vehicle model, vehicle speed and the track quality, each simulation corresponding to 1 km of the line takes around 1 to 10 minutes not including the time needed for post-processing of the results. Therefore, it is basically impossible to perform the simulations for all vehicles on the entire length of all track sections. Instead, we use the load collective method, which is also used in publications such as Enblom (2004) and Dirk and Enblom (2011).

More specifically, this method implies that we create 10 different subsection categories as a function of the track curvature. These subsection categories are track pieces with radii 0-400m, 400-600m, 600-800m, 800-1000m, 1000-1500m, 1500-2000m, 2000-3000m, 3000-5000m, 5000-10000m and above 10000m. Hence, a track section has many track pieces in each subsection category. Considering that we cannot run simulations on the entire track, we choose one track piece in each subsection category (measured by the track geometry car). Specifically, we choose the piece with a track length that is closest to the mean length of all the track pieces in its subsection category. The simulated damage on each piece is then scaled up with respect to the total track length of the subsection category the piece belongs to.

3.5 Simulations results

All four track damage values are calculated for all the subsection categories on each track section and for every vehicle operating on that specific section. Maximum values are considered for all types of damages. The values are then summed for all axles and scaled based on the contribution of the subsection to the entire track section and normalised by the ton-km values obtained from the traffic data.

To show the evaluation process, we present the calculation for a part of track section 217 (see Figure 5). As mentioned earlier, the line is divided into 10 different subsections, depending on the curve radii, and the length of each subsection is presented in Table 2.

Table 2: Subsection lengths of section 217 (based on route length)

Subsection	0-400m	400-600m	600-800m	800-1000m	1000-1500m	1500-2000m	2000-3000m	3000-5000m	5000-10000m	Straight
Total length (m)	0	3543	1584	4368	4139	845	639	565	339	46 627

The traffic data shows that there are eight vehicle categories operating on this line. The corresponding ton-km of each vehicle type is presented in Table 3.

Table 3: Vehicle types & the corresponding ton-km values for section 217

Vehicle types	Million ton-km
Motor coach 4x12t, V_{max} 140 km/h*	0.06
Passenger car 4x14 t, V_{max} 160 km/h	17.89
Motor coach 4x16 t, V_{max} 200 km/h**	56.97
Freight loco 6x20 t, V_{max} 120 km/h	1.63
Freight loco 4x20 t, V_{max} 120 km/h	0.54
Passenger loco 4x19 t, V_{max} 140 km/h	61.45
Passenger loco 4x19 t, V_{max} 175 km/h	19.88
Freight wagon 2x22 t, V_{max} 100 km/h	38.72
Freight wagon 2x6.5 t, V_{max} 100 km/h	12.51
Y25 bogie 4x22 t, V_{max} 100 km/h	659.03
“Unkown”	1.39

*Flexible wheelset guidance, ** Stiff wheelset guidance

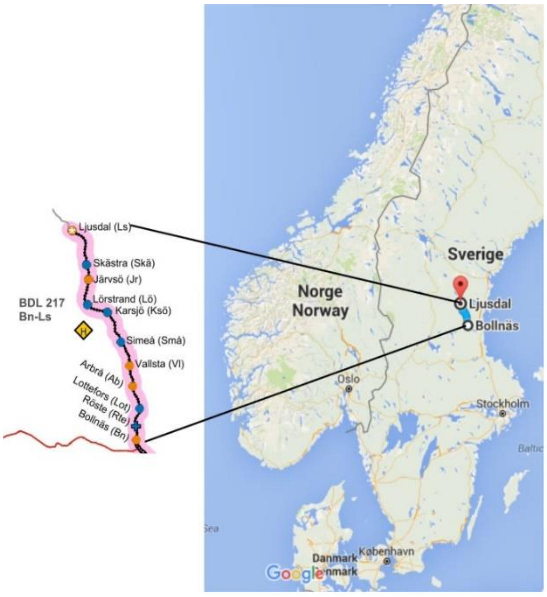


Figure 5: location of section 217

The calculated damages for a “freight loco 4x20 t, V_{max} 120 km/h” running on six segments that constitutes one track piece is presented in Table 4. These measures are based on indices, expect for the damage measure which has the unit J/m. The sum of the maximum wear number for all the axles of the first and the second bogie of this vehicle type, between 83910m to 84510m on section 217, is 1423 J/m. This particular track piece belongs to the curve interval 600-800m, and is scaled accordingly. The total wear on these tracks is: (1423/600)*1584 = 3575 J/m. The same type of calculations are performed for the rest of the curve intervals, including straight lines, in order to produce values of the total wear, RCF, settlement and track component fatigue incurred by this freight loco. Using the weight of the vehicle and the route length of the track section, we calculate its damage values per ton-km. With information on the vehicle’s total ton-km on track section 217, we can scale up the total damage caused by this vehicle on this section. The same type of simulations and calculations are made for rest of the vehicles running on this section to produce measures of total wear, RCF, track settlement, and track component.

Table 4: Results for curve interval 600-800m on section 217 for a freight loco 4x20 t, Vmax 120 km/h

Longitudinal position Start_Stop (m)	Wear (J/m)	RCF (index)	Settlement (index)	Component (index)
83910_84010	134	1.30	5 799 725	3.60E+16
84010_84110	197	2.24	5 740 527	3.52E+16
84110_84210	326	2.96	5 774 048	3.58E+16
84210_84310	322	2.83	5 678 586	3.44E+16
84310_84410	276	2.77	5 679 442	3.43E+16
84410_84510	168	1.84	5 791 845	3.59E+16
Sum	**1423**	**13.94**	**34 464 175**	**2.12E+17**

4. Second stage: Econometric model

With estimates on the damage caused by traffic, we can derive cost elasticities for the damage types using econometric methods. To do so, we need to control for other factors that may influence maintenance costs, such as the average rail age on a section. More specifically, we formulate costs as a function of a set of variables, where the damage types are the variables of main interest

$$C_i = f(D_{1i}, D_{2i}, D_{3i}, D_{4i}, \boldsymbol{X}_i), \qquad (4)$$

where C_i is maintenance costs on $i = 1, 2, \dots, N$ track sections. D_{1i}, D_{2i}, D_{3i}, and D_{4i} are the damage types track settlement, wear of rails, RCF and track component fatigue. $\boldsymbol{X}_i$ is a vector of infrastructure characteristics such as track length and the average age of rails.

As described previously, the damage measures are based on the total ton-km on each section, which in turn depend on the length of each section. Therefore, to separate track length effects from damage effects, we use damage density variables ($\frac{D_{1i}}{Track-km_i}, \frac{D_{2i}}{Track-km_i}$ etc.) along with the track length variable in the model estimations.

In our estimation approach, we start with the translog model proposed by Christensen et al. (1971), which is a second order approximation of a cost (production) function (see for example Christensen and Greene 1976 for an application to cost functions). Both the dependent variable (costs) and the independent variables (damages and infrastructure characteristics) are subject to a logarithmic transformation in this model, which can reduce skewness and heteroscedasticity, problems that may invalidate the statistical inference if not treated correctly. Specifically, we consider A damage types, K network characteristics and M dummy variables, and express the model as

$$\begin{aligned} lnC_i = \alpha + \sum_{a=1}^{A} \beta_a lnD_{ai} + \frac{1}{2}\sum_{a=1}^{A}\sum_{b=a}^{A} \beta_{ab} lnD_{ai}\, lnD_{bi} + \sum_{k=1}^{K} \beta_k lnX_{ki} + \frac{1}{2}\sum_{k=1}^{K}\sum_{l=a}^{K} \beta_{kl} lnX_{ki}\, lnX_{li} \\ + \sum_{a=1}^{A}\sum_{k=1}^{K} \beta_{ak} lnD_{ai} lnX_{ki} + \sum_{m=1}^{M} \beta_m Z_{mi} + v_i \end{aligned} \qquad (5)$$

where α is a scalar, v_i is white noise and $\boldsymbol{\beta}$ is a vector of parameters to be estimated. The simpler (and more restrictive) Cobb-Douglas model is nested in the translog model. We check the Cobb-Douglas constraint $\beta_{ab} = \beta_{kl} = \beta_{ak} = 0$ using an F-test.

5. Data

In total, there were 244 track sections in 2014 administered by the Swedish Transport Administration and their five regional units: Region North, West, East, South and Central. However, limited access to up-to-date track geometry data constrains us to analyze a somewhat smaller part of the Swedish railway network. One may therefore suspect the presence of a selection bias in our data. However, the 143 sections in our data set cover 11 000 track-km out of the 14 100 track-km administered by the Swedish Transport Administration. Hence, the tracks in our data comprise a cross-section of the Swedish rail network with sections from north to south and with large variations in traffic and costs (see Table 5). Still, we can compare the 143 track sections with 169 track sections for which we have information on costs, network characteristics and traffic data. Descriptive statistics of the data are provided in Table 5 (143 sections) and in Table 15 in appendix (169 sections). Estimating a translog cost model generates cost elasticities with respect to ton density at 0.2024 (robust std. error is 0.0479) and 0.2258 (robust std. error is 0.0498), using 143 and 169 track sections, respectively. We therefore consider a (possible) selection bias to be a minor issue in our sample.

Table 5: Descriptive statistics, obs. from 143 track sections

	Median	Mean	St. dev.	Min	Max
Maintenance costs, million SEK	14.25	19.66	17.86	0.87	108.67
Wear	2.21E+12	8.22E+14	5.32E+15	8.26E+06	5.52E+16
RCF	5.58E+08	6.55E+11	3.93E+12	4.46E+05	3.37E+13
Settlement	7.46E+14	5.87E+15	2.75E+16	4.61E+11	2.54E+17
Track component fatigue	3.91E+24	1.85E+28	1.44E+29	4.11E+21	1.38E+30
Wear density	1.09E+08	2.43E+08	4.38E+08	2.10E+06	2.75E+09
RCF density	3.15E+05	5.12E+05	7.46E+05	1.02E+04	7.84E+06
Settlement density	2.96E+12	3.66E+12	3.30E+12	4.89E+10	2.45E+13
Track component fatigue density	2.18E+22	3.37E+22	4.87E+22	3.83E+20	4.80E+23
Route length, km	50.17	60.86	40.59	5.97	215.95
Track length, km	63.95	78.79	52.41	7.84	251.39
Average quality class*	2.77	2.74	1.08	1.00	5.02
Average age of rails	21.2	22.4	9.4	4.1	51.3
Million ton density	4.23	7.68	8.24	0.11	45.72
Region West	0	0.20	0.40	0	1
Region North	0	0.13	0.33	0	1
Region Central	0	0.17	0.38	0	1
Region South	0	0.29	0.45	0	1
Region East	0	0.22	0.42	0	1

* Track quality class ranges from 0-5 (from low to high line speed), but 1 has been added to avoid observations with value 0.

The costs for rectifying track damage are defined as either maintenance or renewal costs. The former are costs for activities conducted to preserve the railway's assets, while the latter are costs for major replacements (minor replacements are defined as maintenance). Given the lumpy nature of renewals, and that we only have access to data for one year (2014), we limit our analysis to maintenance costs only.

Information on the infrastructure characteristics has mainly been collected from the Transport Administration's track information system (BIS), and comprises data on track length, rail age and quality classification (track geometry requirements linked to maximum line speed allowed). As noted in section 3.0, the traffic data contains information on the vehicles (type of wagons, locomotives, multiple unit trains) and their ton-km. The vehicles have been categorized as previously shown in Table 1.

The statistics of the damage measures in Table 5 are based on the total damages incurred by the vehicles on each track section. Table 17 in appendix contains descriptive statistics of the damage measures per vehicle type.

6. Results

Two models are estimated using ordinary least squares (OLS) and the results are presented in Table 6. Model 1 only includes the damage measures, while Model 2 also includes infrastructure characteristics and dummy variables for the regional units, showing the importance of controlling for the production environment in the estimation. All estimations are carried out with Stata 12 (StataCorp.2011).

As a starting point, we examine the correlation coefficients between the different damage mechanisms, which are presented in Table 6. These are all quite high. Track settlement covaries strongly with track component fatigue (the correlation coefficient is 0.95) and with RCF (0.82). The correlation coefficient for wear and track settlement is the lowest (0.72). We therefore also estimate our models using only these two damage mechanisms (*Model 1c*), as we expect them to capture the effects of RCF and track component fatigue to a large extent.

Table 6: Correlation coefficients

	Wear_den.	RCF_den.	Settl. _den.	Comp._den.
Wear_den.	1.0000			
RCF_den.	0.7228	1.0000		
Settl._den.	0.7155	0.8157	1.0000	
Comp._den.	0.8123	0.7752	0.9471	1.0000

As noted in section 4, we start with a full translog model and test linear restrictions of the parameter estimates using F-tests, which results in the restricted translog models presented in Table 7 and 8.

In Model 1, we note that the estimated cost elasticity with respect to track component fatigue is negative (and statistically significant). This result is counterintuitive, indicating that 10 per cent more track component fatigue will *lower* maintenance costs with about 6 per cent. However, the variance inflation factors (VIFs) for the first order coefficients for settlement and track component fatigue are 0.17 and 0.20, respectively. Also, considering the high correlation coefficients, we drop track component fatigue, which results in *Model 1b*. The first order coefficient for settlement then falls from 0.62 to -0.03 (not significantly different from zero). Dropping RCF due to its high correlation with settlement (0.82), results in Model 1c. The sum of the first order coefficients are rather similar in the models (0.44, 0.33 and 0.40), which indicates that the strong correlation between the damages mechanisms affects the individual parameter estimates significantly.

Table 7: Estimation results, Model 1

	Model 1a		*Model 1b*		*Model 1c*	
	Coef.	Rob. Std. Err.	Coef.	Rob. Std. Err.	Coef.	Rob. Std. Err.
Cons.	16.5063***	0.1124	16.3996***	0.0763	16.5134***	0.0826
Wear_den.	0.3616**	0.1495	0.0137	0.0974	0.2845**	0.1151
Wear_den.^2	-0.4863**	0.2443	-	-	-0.3613***	0.1183
RCF_den.	0.0885	0.1520	0.3544***	0.1195	-	-
Settl. _den.	0.6205**	0.3091	-0.0359	0.1230	0.1178	0.1080
Comp. _den.	-0.6328**	0.2965	-	-	-	-
Comp. _den.^2	-0.1029	0.2961	-	-	-	-
Wear_den.Settl. _den.	-	-	-	-	0.2189**	0.0929
Wear_den.Comp. _den.	0.3396	0.2479	-	-	-	-
Mean VIF	15.58		3.01		3.06	
R^2	0.22		0.14		0.16	
Adj. R^2	0.18		0.12		0.14	

We transform all data by dividing by the sample median prior to taking logs. In that way, the first order coefficients can be interpreted as cost elasticities at the sample median. See Table 13 in appendix for definitions of the variables.

Note: ***, **, *: Significance at 1 %, 5 %, and 10 % level, respectively

Leaving out cost drivers that are correlated with the damage measures may lead to undesirable omitted variable bias. If that is the case, the coefficients in Model 1 are over- or underestimated. Thus, in Model 2, we include a set of control variables that we believe to be important in this context. Here it should be noted that differences in track irregularities, curvature, line speeds and traffic volume have been (at least substantially) normalized, as these aspects are inputs in the simulations and therefore picked up by the damage measures. However, including Qual_ave (linked to line speed) can pick up quality aspects other than wear and tear caused by vehicles' line speed, for example maintenance strategies/priorities associated with line speed – that is, there is a difference between the damage caused by different track qualities (picked up by the damage measures) and the cost of correcting this damage with respect to quality class (due to strategies/priorities). We also include rail age to control for maintenance costs that are due to previous use of the track rather than the damage caused by traffic in 2014. A track length variable is included as we use damage density variables ($\frac{D_{1i}}{Track-km_i}, \frac{D_{2i}}{Track-km_i}$ etc.) in the model to separate track length effects from damage effects. However, we do not expect track length will pick up scale effects *per se* in this study (groups of track sections belong to contract areas), but is included as the length of a section may lack a one-to-one relationship with other infrastructure characteristics. For example, the correlation coefficient with switch length is 0.66 in a subset of our data (preferably, a switch length variable would have been included, but was not available for all the 143 track sections). Moreover, we include dummy variables for maintenance regions, as these may pick up differences in the management of the sections.

Higher speeds imply stricter requirements on track quality (track geometry). This may increase the propensity to rectify the settlement damage caused by the vehicles. Indeed, the interaction term between Settlement and Qual_ave is negative, which suggests that the cost impact of settlement is lower for low line speeds compared to high line speeds. The first order coefficient for Qual_ave is negative, yet not significant. We calculate the cost elasticities with respect to Qual_ave at the observed levels of the variables (which in Model 2 is $\hat{\gamma}_{iQual.} = \hat{\beta}_5 + 2 \cdot \hat{\beta}_6 lnQualave_i + \hat{\beta}_8 lnSettlement_i$), which shows that the elasticities are positive for low levels of Qual_ave (high linespeed) and turn negative for high levels of the variable (low linespeed).

Table 8: Estimation results, Model 2

	Model 2a		*Model 2b*	
	Coef.	Rob. Std. Err.	Coef.	Rob. Std. Err.
Cons.	16.4675***	0.1030	16.4779***	0.1024
Wear_den.	0.1079	0.0718	0.1182*	0.0714
RCF_den.	0.0485	0.0805	-	-
Settl._den.	0.0996	0.0983	0.1345*	0.0719
Track_length	0.9303***	0.0582	0.9385***	0.0588
Qual_ave	-0.0428	0.2185	-0.0237	0.2113
Qual_ave^2	-0.9850*	0.5132	-1.0099*	0.5205
Rail_age	0.2575*	0.1337	0.2699**	0.1340
Settl._den.Qual_ave	-0.5618***	0.1189	-0.5685***	0.1188
Region_West	0.3264**	0.1429	0.3254**	0.1431
Region_North	0.0442	0.1903	0.0395	0.1886
Region _Central	-0.2981**	0.1500	-0.2933*	0.1491
Region _South	-0.2179	0.1395	-0.2173	0.1395
Mean VIF	2.73		2.31	
R^2	0.70		0.70	
Adj. R^2	0.67		0.67	

We transform all data by dividing by the sample median prior to taking logs. In that way, the first order coefficients can be interpreted as cost elasticities at the sample median. See Table 13 in appendix for definitions of the variables.

Note: ***, **, *: Significance at 1 %, 5 %, and 10 % level, respectively

Older rails seem to be more costly. Considering that rail age is a proxy for track standard due to accumulated use, a positive and significant coefficient is intuitive as high maintenance costs on old and heavily used track is expected, which eventually makes a renewal economically justified. The track length coefficient is close to 1 in the model estimations, indicating that there is no scale economy with respect to this variable (we cannot reject the null hypothesis of constant economies of scale, F(1, 131)=1.09, prob>F=0.2977).

Turning to the cost elasticities with respect to the damage measures in Model 2a, we note that these are 0.1079, 0.0485 and 0.0996 for wear, RCF and settlement, respectively. None of these estimates are statistically significant. In Model 2b we drop RCF due to its high correlation coefficient with settlement, which generates a slightly higher estimate for settlement. The coefficients for wear and settlement are now statistically significant at the 10 per cent level. The sum of the first order coefficients is 0.2560 and 0.2527 in Model 2a and Model 2b, respectively, indicating that the cost impact of RCF is to a large extent picked up by the estimates for wear and settlement.

6.1 Marginal costs

To calculate the marginal costs of different vehicle types, we first need to calculate the marginal cost per damage unit (see section 6.1.1 below). Similar to Smith et al. (2017), we then link these costs to vehicle types based on the amount of damage per ton-km each vehicle has caused according to the simulations in the first stage of our estimation approach. In that way, we produce a marginal cost per ton-km, which is the preferred charging unit (see section 6.1.2 and section 6.1.3 below).

6.1.1 Marginal cost per damage unit

In the marginal cost estimation presented below, we use the estimated cost elasticities for wear and settlement (evaluated at the sample median). Marginal costs that are based on the non-significant cost elasticities in Model 2a are presented in Table 14 in appendix.

The marginal cost of a damage mechanism j is formulated as

$$MC_{ij}\ per\ damage\ unit = \frac{\partial C_i}{\partial D_{ij}} = \frac{D_{ij}}{C_i}\frac{\partial C_i}{\partial D_{ij}}\frac{C_i}{D_{ij}} = \frac{\partial lnC_i}{\partial lnD_{ij}}\frac{C_i}{D_{ij}}, \quad (6)$$

where D is damage. Hence, from equation (6) we can express the marginal cost estimate at track section i as

$$MC_{ij}\ per\ damage\ unit = \hat{\gamma}_j \cdot \widehat{AC}_{ij}, \quad (7)$$

where $\hat{\gamma}_j$ is the estimated cost elasticity ($\frac{\partial lnC_i}{\partial lnD_{ij}}$) of damage mechanism j. $\widehat{AC}_{ij}$ is the average cost ($\frac{\hat{C}_i}{D_{ij}}$), where $\hat{C}_i$ is predicted costs specified as

$$\hat{C}_i = \exp[\ln(C_i) - \hat{v}_i + 0.5\hat{\sigma}^2], \quad (8)$$

Equation (8) derives from the double-log specification and the assumption of normally distributed residuals (see for example Munduch et al. 2002).

We use a weighted marginal cost for the 143 track sections in this study, according to equation (9) below. Specifically, we use each track section's share of total damage as weights and multiply with each section's marginal cost per damage unit. Taking the sum over all track sections produces a weighted marginal cost estimate that generates the same income as if each section's marginal cost would be used.

$$MC_j^W \, per\ damage\ unit = \sum_i \left[MC_{ij} \, per\ damage\ unit \cdot \frac{D_{ij}}{\sum_i D_{ij}} \right], \qquad (9)$$

The average cost, the unweighted and weighted marginal costs are presented in Table 9. These costs become quite low as they are estimates per total damage.

Table 9: Average and marginal costs per damage unit, SEK in 2014 prices

	Variable	Mean	Std. Err.	[95% Conf.	Interval]
Average cost	Wear	5.62E-03	7.90E-04	4.05E-03	7.18E-03
	Settlement	2.56E-07	5.08E-08	1.56E-07	3.57E-07
Unweighted marginal cost	Wear	6.64E-04	9.34E-05	4.79E-04	8.48E-04
	Settlement	3.45E-08	6.84E-09	2.10E-08	4.80E-08
Weighted marginal cost	Wear	1.41E-04	-	-	-
	Settlement	9.37E-09	-	-	-

The marginal cost for settlement is lower than the cost for wear, even though their respective cost elasticities are similar. The reason is that the damages have different units, generating an average cost of settlement that is much lower than the average cost of wear.

6.1.2 Average marginal cost per ton-km

Estimates that are comparable with the costs in the literature on marginal rail infrastructure costs in Sweden (and elsewhere) are produced when multiplying the estimates in Table 9 with the damage caused by a ton-km of a certain vehicle.[3] However, before we estimate a marginal cost per ton-km for each vehicle type (see section 6.1.3), we examine the average marginal costs per ton-km for all vehicles to make a comparison with previous estimates on Swedish data. To do this, we need to multiply the marginal cost per damage type with the damage per ton-km on each track section, which can be either an average or a weighted average damage per ton-km caused by the vehicles.

[3] In fact, this calculation normalizes the differences in units between the damage mechanism; see Table 11, where the differences in MC per ton-km between wear and settlement do not differ at the same order of magnitude as the marginal costs in Table 9.

Other alternatives are to use an average or weighted average damage per ton-km over all track sections.

The first set of weighted averages are calculated by using the vehicle types' (v) shares of total gross-ton km on each track section i as weights, and then by multiplying these weights with the damage per ton-km values for wear and settlement. Taking the sum over all vehicle types v for each track section i and damage type j, generates weighted averages of wear per ton-km and settlement per ton-km, which is then multiplied with the marginal cost per damage unit (see equations 10 and 12).

To get weighted averages per ton-km over all track sections, we use shares of the total gross ton-km (sum of all 143 track sections) as weights, multiply the weights with the damage per ton-km values, and take the sum over all vehicle types v and track sections i for each damage type j (see equations 11 and 13).

Note that we can use either the unweighted marginal cost per damage unit (equation 7) or the weighted marginal cost per damage unit (equation 9) in these calculations; as mentioned in the previous sections, using the latter will produce a marginal cost estimate that generates the same income as if each section's marginal cost would be used. Specifically, to calculate the unweighted marginal cost per ton-km for track section i and damage type j, we use the unweighted marginal cost per damage unit (equation 7) and the weighted average damages per ton-km for each track section

$$MC_{ij}\ per\ tonkm = MC_{ij}\ per\ damage\ unit \cdot \sum_v \left[\frac{D_{ijv}}{GTkm_{iv}} \cdot \frac{GTkm_{iv}}{(\sum_v GTkm_{iv})}\right], \qquad (10)$$

or the weighted average damage per ton-km over all track sections

$$MC_{ij}\ per\ tonkm = MC_{ij}\ per\ damage\ unit \cdot \sum_i \sum_v \left[\frac{D_{ijv}}{GTkm_{iv}} \cdot \frac{GTkm_{iv}}{(\sum_i \sum_v GTkm_{iv})}\right], \qquad (11)$$

To calculate the weighted marginal cost per ton-km for damage type j, we use the weighted marginal cost per damage unit (equation 9) and multiply with either weighted average damages per ton-km for each track section or all track sections (equation 12 and 13, respectively)

$$MC_j^W per\ tonkm = MC_j^W per\ damage\ unit \cdot \sum_v \left[\frac{D_{ijv}}{GTkm_{iv}} \cdot \frac{GTkm_{iv}}{(\sum_v GTkm_{iv})}\right], \qquad (12)$$

$$MC_j^W per\ tonkm = MC_j^W per\ damage\ unit \cdot \sum_i \sum_v \left[\frac{D_{ijv}}{GTkm_{iv}} \cdot \frac{GTkm_{iv}}{(\sum_i \sum_v GTkm_{iv})}\right], \qquad (13)$$

These marginal costs are summarized in Table 10 below.

Table 10: Average unweighted and weighted marginal costs per ton-km, SEK

Equation	Damage per ton-km	Marginal cost (MC)	Damage				
				Mean	Std. Err.	[95% Conf.	Interval]
Eq. 10	Weighted averages for each section	Unweighted MC	Wear	0.0246	0.0049	0.0148	0.0344
			Settlement	0.0268	0.0053	0.0164	0.0373
			Total	0.0514	0.0102	0.0313	0.0716
Eq. 11	Weighted average for all sections		Wear	0.0293	0.0041	0.0211	0.0374
			Settlement	0.0259	0.0051	0.0158	0.0361
			Total	0.0552	0.0082	0.0389	0.0715
Eq. 12	Weighted averages for each section	Weighted MC	Wear	0.0085	0.0013	0.0059	0.0112
			Settlement	0.0072	0.0005	0.0063	0.0081
			Total	0.0157	0.0015	0.0127	0.0187
Eq. 13	Weighted average for all sections		Wear	0.0062	0.0004	0.0054	0.0070
			Settlement	0.0070	0.0004	0.0062	0.0079
			Total	0.0132	0.0008	0.0116	0.0149

The unweighted marginal costs per ton-km for wear and settlement (equation 11) are illustrated in Figure 6, showing that costs fall sharply with ton density. Similar shapes were found for a number of European countries (including Sweden) in Wheat et al. (2009).

Using weighted averages of damage per ton-km for each track section generates a weighted marginal cost per ton-km for wear at 0.0085 SEK and 0.0072 SEK for settlement (which in total is 0.0157 SEK). The corresponding costs are somewhat lower when using a weighted average over all track sections: it is 0.0062 SEK for wear, 0.0070 SEK for settlement, and 0.0132 SEK in total. These estimates are higher than previous estimates on Swedish data in Andersson (2008) and Odolinski and Nilsson (2017), which are 0.0080 SEK and 0.0094 SEK respectively (in 2014 prices).

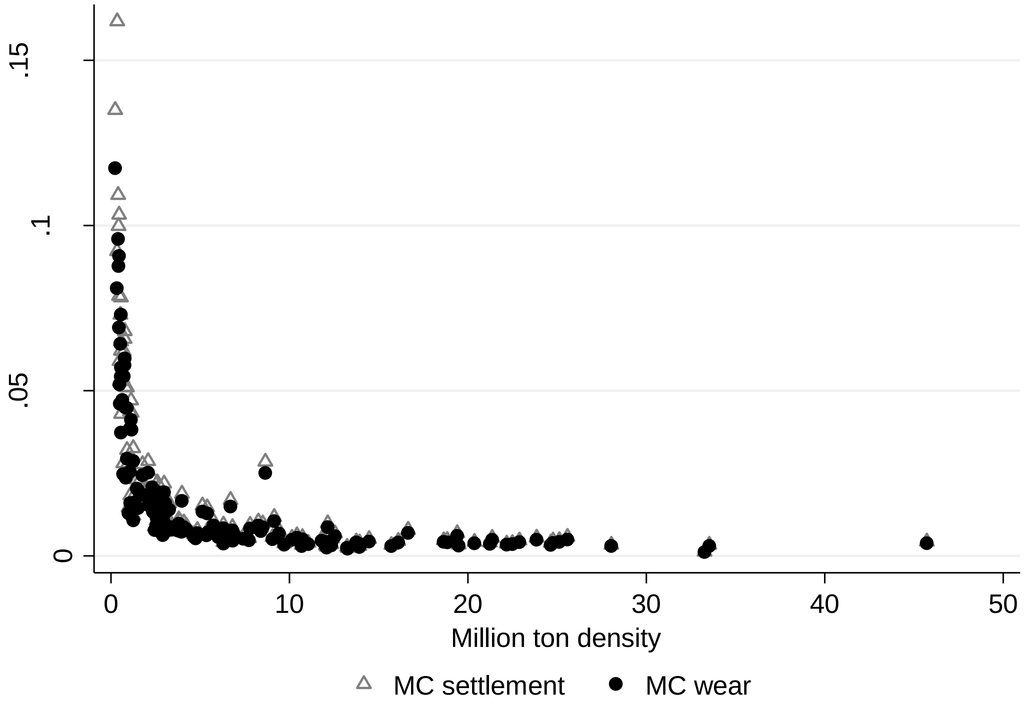


Figure 6: Unweighted marginal costs for settlement and wear (eq. 11)

6.1.3 Marginal cost per ton-km and vehicle type

To calculate a weighted marginal cost per ton-km for each vehicle type, we need to use a different set of weights compared to equations 10-13 (see equation 14). Specifically, these weights are calculated using the sum of ton-km over all track sections for each vehicle type, i.e. the weights are a vehicle type's share of gross ton-km on track section i with respect to the vehicle type's total gross ton-km. Using the product of the weights and the damage values for wear and settlement, and taking the sum over all track sections, we get the weighted average damage j for each vehicle type v (see Table 11; descriptive statistics of all damage measure per vehicle type are presented in Table 17 in appendix). Multiplying these weighted averages with the weighted marginal cost per damage unit generates a weighted marginal cost per ton-km for vehicle type v and damage j. For example, a freight loco 4x20 t, V$_{max}$ 120 km/h, has a weighted average wear per ton-km at 21.75 and a weighted average settlement per ton-km at 743 656. Its total marginal cost per ton-km is therefore 21.75*1.41-E04 + 743 656*9.37E-09 = 0.0100 SEK, where 1.41-E04 and 9.37E-09 are the weighted marginal cost for wear and settlement, respectively.

$$MC_{jv}^{W} per\ tonkm = MC_{j}^{W} per\ damage\ unit \cdot \sum_i \left[D_{ijv} \cdot \frac{GTkm_{iv}}{(\sum_i GTkm_{iv})}\right], \qquad (14)$$

The marginal costs per vehicle and damage type are presented in Table 11 (the damage measures for RCF are presented in Table 16 in appendix). Note that the marginal costs in Table 11 partly depend on which track sections the different vehicles ran on during 2014, where track quality differs between the sections. More specifically, we used measurements on track geometry and track irregularities as input in the simulation, as this will affect the damage caused by traffic (the resulting total damage values are essential in the estimation of the cost elasticities). Hence, considering that each vehicle type did not run on all the 143 track sections, the values in Table 11 are not completely normalized.[4]

Table 11: Damages and marginal costs (MC_{jv}^{W}) per ton-km and vehicle type

Vehicle type	Wear per ton-km	Settlement per ton-km	MC wear[a]	MC settlement [a]	Total MC[a]
Motor coach 4x21 t, V_{max} 200 km/h *	209.76	995 468	0.0295	0.0093	0.0389
Three-piece bogie 4x30 t, V_{max} 60 km/h	97.56	867 067	0.0137	0.0081	0.0219
Passenger car 4x14 t, V_{max} 160 km/h	57.34	741 423	0.0081	0.0069	0.0150
Freight loco 6x30 t, V_{max} 70 km/h	36.85	1 001 992	0.0052	0.0094	0.0146
Freight loco 6x20 t, V_{max} 120 km/h	36.90	945 300	0.0052	0.0089	0.0141
Motor coach 4x16 t, V_{max} 200 km/h**	41.46	852 697	0.0058	0.0080	0.0138
Passenger Loco 4x19 t, V_{max} 175 km/h	40.69	740 151	0.0058	0.0070	0.0128
Three-piece bogie 4x6.5 t, V_{max} 60 km/h	50.22	602 992	0.0071	0.0056	0.0127
Passenger Loco 4x19 t, V_{max} 140 km/h	40.85	748 934	0.0057	0.0069	0.0127
Motor coach, Jacob bogie 3x16.5 t, V_{max} 160 km/h**	53.58	476 803	0.0075	0.0045	0.0120
Y25 bogie 4x22 t, V_{max} 100 km/h	30.32	795 901	0.0043	0.0075	0.0117
Freight wagon 2x6.5, V_{max} 100 km/h	49.75	383 151	0.0070	0.0036	0.0106
Motor coach, Jacob bogie 3x12.5 t, V_{max} 200 km/h***	33.73	571 887	0.0048	0.0054	0.0101
Freight loco 4x20 t, V_{max} 120 km/h	21.75	743 656	0.0031	0.0070	0.0100
Motor coach 4x12 t, V_{max} 140 km/h***	21.12	668 032	0.0030	0.0063	0.0092
Freight wagon 2x22 t, V_{max} 100 km/h	26.48	464 017	0.0037	0.0043	0.0081
Motor coach 4x16 t, V_{max} 200 km/h***	12.03	676 894	0.0017	0.0063	0.0080
All vehicles, weighted average (eq. 13)	44.10	751 142	0.0062	0.0700	0.0132

[a] SEK in 2014 prices * High center of gravity and stiff wheelset guidance, ** Stiff wheelset guidance, ***Flexible wheelset guidance

The vehicles in Table 11 are ordered after the highest marginal cost, showing that Motor coach 4x21 t, V_{max} 200 km/h (stiff wheelset guidance and high center of gravity) is assigned a marginal cost at 0.0389 SEK. The other estimates stretch from 0.0080 SEK to 0.0219 SEK, indicating rather differentiated marginal costs. Interestingly, a tare freight wagon 2x22t, V_{max} 100km, has a higher marginal cost (0.0106 SEK) than its laden counterpart, which has a marginal cost at 0.0081 SEK. The reason for this relationship is that the tare freight wagon has a factor 1.88 higher wear per ton-km

[4] Normalized marginal costs can be generated by using damage values for each vehicle type on a representative track in equation (14), i.e. the only difference in damages per ton-km would be caused by the vehicle type's characteristics and speed.

(weighted average) than the laden freight wagon, while the laden wagon only has a factor 1.21 higher settlement per ton-km than the tare wagon (cf. Table 11). Considering that the cost elasticities for the different damage types are rather similar, these differences in damages are reflected in the marginal costs.

The marginal maintenance costs in this paper differ from the costs generated by Öberg et al. (2007, p. 58-59). The relative differences between vehicle types in their study does not match the relative differences reported in Table 11 (or Table 16). One reason is that they use cost shares for the damages mechanisms reported by the Swedish Rail Administration; settlement is responsible for 25 per cent, wear and RCF was attributed 40 per cent, while component fatigue was allocated 35 per cent of costs. To make a comparison between these cost shares and our estimates, we normalize the sum of our cost elasticities to 1, and calculate their respective shares. For example, the cost elasticity with respect to wear in Model 2a is 0.1079, while it is 0.0996 for settlement and 0.0485 for RCF, which corresponds to a cost share at (0.1079/(0. 1079+0.0996+0.0485) = 0.42 for wear, (0.0996/(1079+0.0996+0.0485)) = 0.39 for settlement and (0.0485/(1079+0.0996+0.0485)) = 0.19 for RCF. We also estimate these cost shares for Model 2b and summarize in Table 12 below. Note that, due to the high correlation coefficients (presented in Table 6), the settlement estimates are considered to capture the cost impact from track component fatigue in our model estimations, while in Model 2b, both the wear and settlement estimates are considered to capture the cost impact of RCF.

Table 12: Cost shares of damage mechanisms

Damage	Model 2a	Öberg et al. 2007	Damage	Model 2b
Wear + RCF	0.61	0.4	Wear + RCF*	0.47
Settl. + track comp. fatigue	0.39	0.6	Settl. + RCF*+ track comp. fatigue	0.53

* share of RCF

The cost shares from Model 2a are the mirror image of the shares in Öberg et al., while the cost shares from Model 2b are somewhat closer (yet, the latter comparison is shaky as the cost impact from RCF is captured by both the wear and settlement estimate in Model 2b). There are, however, significant differences in the estimation approaches. Öberg et al. perform their simulations on a "representative" track with a curve distribution that was weighted by the actual traffic volume on different curve zones on 5000 km of tracks, which is about 35 per cent of the total network length. Added to this, their simulations were carried out on perfect tracks with track gauge 1435 (no track irregularities), except for freight vehicles, which were simulated on a track with irregularities based on measurements on a 500 m section of the Swedish main line. Hence, both the simulation strategy and the cost calculations in Öberg et al. (2007) differ from our paper, as we perform simulations based on curvature and track measurements of irregularities on each of the 143 track sections (comprising almost 80 per cent of the total network length), in order to predict the actual damage from traffic during a year and relate it to actual costs during the same year. Moreover, as noted in the introduction and in the methodology section (2), a significant contribution of our paper is that we estimate the current cost shares attributed to the different damage mechanisms.

Finally, it should be pointed out that the cost elasticities we use in the marginal cost estimation are considered to also capture effects of RCF and track component fatigue. Indeed, this seems to be the case (at least for RCF) as the sum of the average weighted MC for all damage types (equation 13) from Model 2a is 0.0134 SEK, which is practically the same estimate generated by Model 2b (0.0132 SEK). However, the correlation between the vehicle's damages per ton-km (weighted averages, cf.

equation 14) is quite low compared to the correlation coefficients in Table 6 that are calculated for track sections.[5] For example, the correlation coefficient between the weighted averages of wear and RCF for the different vehicles is 0.21, and 0.53 between settlement and track component fatigue (however, the correlation coefficient wear and component fatigue is 0.82). This implies that the relationship between the vehicles in Table 11 would be different if we had been able to isolate the relative costs of all damage mechanisms. An indication of this are the marginal costs for each vehicle type in Model 2a (presented in Table 16 in appendix), which include the cost impact from RCF. This generates a slightly different relationship between the vehicles' costs, which can be summarized by the difference in the rankings of the vehicle types with respect to their total marginal cost (cf. Table 16; the correlation coefficient between the different rankings is 0.74).[6] Hence, the estimates presented in Table 11 should be interpreted with care.

7. Conclusion

This paper contributes to the existing literature by showing that the two-stage method in Smith et al. (2017) can produce estimates on the relative cost of damage mechanisms that can be informative for infrastructure managers in Europe. Specifically, by combining engineering and econometric approaches, we have estimated marginal costs for the vehicle types running on the Swedish railway network. We have developed previous work on this method by using a larger set of - and more detailed - vehicle models, as well as a larger set of track sections that constitutes a major part of the Swedish railway network.

The different damage mechanisms proved to be highly correlated between track sections, making it difficult to isolate the cost impact of each damage type. Still, our model was able to provide significant cost elasticities with respect to wear and settlement that could be used in the estimation of marginal costs. The estimates for these two damage mechanisms capture the cost impact from RCF and track component fatigue to a large extent. However, the downside is that the marginal costs do not reflect the relative differences in RCF per ton-km and track component fatigue per ton-km between the vehicle types.

The results in this paper indicate a substantial variation in the marginal cost per ton-km for different vehicle types running on the Swedish railway, which is due to differences in the damage done by the vehicles and the relative cost of the damage mechanisms. Track access charges with similar relative differences between vehicle types would create strong incentives for using more track friendly vehicles.

More observations over time can be valuable for future research to generate more reliable and robust estimates. The results from our approach can also be used to differentiate track access charges with respect to, for example, line speed, in line with the charges that Switzerland has proposed to implement in 2017. More specifically, future work can use the simulation results on how line speed adds to different damages. Together with the relative costs of these damage mechanisms,

[5] A track section with poor track geometry and many (sharp) curves is likely to have a high wear, settlement, RCF and track component fatigue (*ceteris paribus*), and vice versa. This can explain the high correlation coefficients between the damage mechanisms at the track section level, even though the correlation coefficients between the damages done by different vehicle types are low.

[6] However, note that costs in Table 16 are based on cost elasticities with respect to damages that are not statistically significant.

it is then possible to calculate marginal costs for different vehicles that are also differentiated with respect to line speeds.

8. References

Andersson, M. (2007): 'Empirical essays on railway infrastructure costs in Sweden', Doctoral Thesis in Infrastructure with specialization in Transport and Location Analysis, June 2007. Department of Transport Economics, KTH, Stockholm, ISBN 10: 91-85539-18-X.

Andersson, M. (2008): 'Marginal Railway Infrastructure Costs in a Dynamic Context', *EJTIR*, 8, 268-286.

Andersson, E., M. Berg, and S. Stichel (2015): 'Rail Vehicle Dynamics', Division of Railway Technology, Department of Aeronautical and Vehicle Engineering, Royal Institute of Technology (KTH), Stockholm.

Banverket (1996): 'Tillåten hastighet m h t spårets geometriska form', BVF 586.41 (In Swedish).

Booz Allen Hamilton, and TTCI (2005): 'Review of Variable Usage and Electrification Asset Usage Charges: Final Report', Report R00736a15, London, June 2005.

Burstow M.C. (2004): 'Whole life rail model application and development for RSSB continued development of RCF damage parameter', AEATR-ES-2004-880, Issue 2, Rail Safety & Standards Board (RSSB).

CEN (2010): 'Railway applications – Structural Requirements of Railway Vehicle Bodies', EN 12663.

Chaar, N. and M. Berg (2006): 'Simulation of vehicle track interaction with flexible wheelsets, moving track models and field tests', *Vehicle System Dynamics* 44 (1), 921-931.

Christensen, L.R and W.H. Greene (1976): 'Economies of Scale in U.S. Electric Power Generation', *The Journal of Political Economy,* 84(4), 655-676.

Christensen, L.R, D.W. Jorgenson, and L.J. Lau (1971): 'Transcendental Logarithmic Production Frontiers', *The Review of Economics and Statistics*, 55(1), 28-45.

Cohen, J., P. Cohen, S. G. West and L. S. Aiken (2003): 'Multiple Regression/Correlation Analysis for the Behavioral Sciences', Third Edition, Lawrence Erlbaum Associates, Inc.

Enblom, R. (2004): 'Simulation of wheel and rail profile evolution-wear modelling and validation'. TRITA AVE 2004:19, Licentiate Thesis, Department of Aeronautical and Vehicle Engineering, Royal Institute of Technology (KTH), Stockholm.

Dirk, B., and R. Enblom (2011): 'Prediction model for wheel profile wear and rolling contact fatigue', *Wear*, 271 (1-2), 210-217.

GENSYS 15.04 (2015): Online manual www.gensys.se; AB Desolver, Östersund.

Johansson, P., and J-E. Nilsson (2004): 'An economic analysis of track maintenance costs', *Transport Policy*, 11, 277-286.

Munduch, G., A. Pfister, L. Sögner, and A. Siassny (2002): 'Estimating Marginal Costs for the Austrian Railway System', *Vienna University of Economics & B.A., Working Paper* No. 78, February 2002.

Link, H. A. Stuhlehemmer, M. Haraldsson, P. Abrantes, P. Wheat, S. Iwnicki, C. Nash, and A.S.J Smith (2008): 'CATRIN (Cost Allocation of TRansport INfrastructure cost), Deliverable D 1, Cost allocation Practices in European Transport Sector. VTI, Stockholm, March 2008.

ORR (2013): 'Periodic Review 2013: Final determination of Network Rail's outputs and funding for 2014-2019', October 2013, Office of Rail Regulation (ORR).

Odolinski, K. and J-E. Nilsson (2017): 'Estimating the marginal cost of rail infrastructure maintenance using static and dynamic model; does more data make a difference?', Economics of Transportation, 10, 8-17.

Smith, A. S. J., Iwnicki, S., Kaushal, A. Odolinski, K., and Wheat, P. (2017): 'Estimating the relative cost of track damage mechanisms: combining economic and engineering approaches', *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, (In press)

StataCorp.2011. *Stata Statistical Software: Release 12*, College Station, TX: StataCorp LP.

UIC/ORE Question D 161.1 (1987): 'Dynamic effect of 22.5 t axle loads on the track', Report no. 4.

Wheat, P., A.S.J. Smith, and C. Nash (2009): 'CATRIN (Cost Allocation of TRansport INfrastructure cost)', Deliverable 8 – Rail Cost Allocation for Europe, VTI, Stockholm.

Öberg, J., E. Andersson, and J. Gunnarsson (2007): 'Track Access Charging with Respect to Vehicle Characteristics', second edition, Rapport LA-BAN 2007/31

9. Appendix

Table 13: Definition of variables

Wear_den.	=	ln(wear density)
Wear_den.^2	=	ln(wear density)*ln(wear density)
RCF_den	=	ln(RCF density)
Settl._den	=	ln(Settlement density)
Comp._den,	=	ln(Track component fatigue density)
Comp._den.^2	=	ln(Track component fatigue density)*ln(Track component fatigue density)
Wear_den.Settl._den	=	ln(wear density)*ln(settlement density)
Wear_den.Com._den.	=	ln(wear density)*ln(Track component fatigue density)
Track_length	=	ln(track length)
Qual_ave	=	ln(average quality class)
Qual_ave^2	=	ln(average quality class)* ln(average quality class)
Rail_age	=	ln(average age of rails)
Settl._den.Qual_ave	=	ln(settlement density)*ln(average quality class)
Region_West	=	Dummy for region West
Region_North	=	Dummy for region North
Region_Central	=	Dummy for region Central
Region_South	=	Dummy for region South
Region_East	=	Dummy for region East

Table 14: Model 2a - Average and marginal costs per damage unit, SEK in 2014 prices

	Variable	Mean	Std. Err.	[95% Conf.	Interval]
Average cost	Wear	5.60E-03	7.84E-04	4.05E-03	7.15E-03
	RCF	1.65E+00	2.12E-01	1.23E+00	2.07E+00
	Settlement	2.58E-07	5.27E-08	1.54E-07	3.63E-07
Marginal cost	Wear	6.04E-04	8.46E-05	4.37E-04	7.72E-04
	RCF	8.02E-02	1.03E-02	5.99E-02	1.01E-01
	Settlement	2.57E-08	5.25E-09	1.53E-08	3.61E-08
Weighted marginal cost	Wear	1.29E-04	8.21E-06	1.13E-04	1.45E-04
	RCF	2.34E-02	1.49E-03	2.04E-02	2.63E-02
	Settlement	6.94E-09	4.43E-10	6.07E-09	7.82E-09

Table 15: Descriptive statistics, obs. from 169 track sections

	Median	Mean	St. dev.	Min	Max
Maintenance costs, million SEK	13.71	19.71	22.42	0.53	209.22
Track length, km	58.71	72.51	51.85	4.52	251.39
Average quality class	2.89	2.94	1.15	1.00	5.17
Average age of rails	21.2	22.3	9.6	2.3	53.1
Million ton density	4.52	8.15	9.39	0.00	61.98
Region West	0	0.20	0.40	0	1
Region North	0	0.14	0.34	0	1
Region Central	0	0.18	0.39	0	1
Region South	0	0.25	0.43	0	1
Region East	0	0.24	0.43	0	1

Table 16: Model 2a - Marginal costs per ton-km and vehicle type (based on non-statistically significant cost elasticities)

Vehicle type	Wear[a]	Settlement[a]	RCF[a]	Total MC[b]	Rank #	Rank # Model 2b
Motor coach 4x21 t, V_{max} 200 km/h*,**	209.76	995 468	0.09	0.0361	1	1
Three-piece bogie 4x30 t, V_{max} 60 km/h	97.56	867 067	0.26	0.0247	2	2
Freight wagon 2x6.5, V_{max} 100 km/h	49.75	383 151	0.34	0.0169	3	12
Passenger car 4x14 t, V_{max} 160 km/h	57.34	741 423	0.14	0.0159	4	3
Freight loco 6x20 t, V_{max} 120 km/h	36.90	945 300	0.11	0.0139	5	5
Motor coach 4x16 t, V_{max} 200 km/h**	41.46	852 697	0.11	0.0138	6	6
Three-piece bogie 4x6.5 t, V_{max} 60 km/h	50.22	602 992	0.12	0.0134	7	8
Motor coach, Jacob bogie 3x16.5 t, V_{max} 160 km/h**	53.58	476 803	0.13	0.0132	8	10
Freight loco 6x30 t, V_{max} 70 km/h	36.85	1 001 992	0.05	0.0128	9	4
Passenger Loco 4x19 t, V_{max} 175 km/h	40.69	740 151	0.07	0.0127	10	7
Passenger Loco 4x19 t, V_{max} 140 km/h	40.85	748 934	0.10	0.0121	11	9
Motor coach, Jacob bogie 3x12.5 t, V_{max} 200 km/h***	33.73	571 887	0.14	0.0116	12	13
Y25 bogie 4x22 t, V_{max} 100 km/h	30.32	795 901	0.08	0.0114	13	11
Freight loco 4x20 t, V_{max} 120 km/h	21.75	743 656	0.07	0.0097	14	5
Freight wagon 2x22 t, V_{max} 100 km/h	26.48	464 017	0.09	0.0088	15	16
Motor coach 4x12 t, V_{max} 140 km/h***	21.12	668 032	0.05	0.0085	16	15
Motor coach 4x16 t, V_{max} 200 km/h***	12.03	676 894	0.05	0.0073	17	17
All vehicles, weighted average (eq. 13)	44.10	751 142	0.11	0.0134		

[a] per ton-km [b] SEK in 2014 prices * High center of gravity, ** Stiff wheelset guidance, ***Flexible wheelset guidance

Table 17: Descriptive statistics of damage measures per vehicle type

	Wear				RCF				Settlement				Component			
Vehicle	Mean	St.dev	Min	Max	Mean	St.dev	Min	Max	Mean	St.dev	Min	Max	Mean	St.dev	Min	Max
Freight loco 4x20 t, 120 km/h	32.4	106.2	0.9	951	0.06	0.07	0.00	0.47	743 564	79 808	701 964	1 290 819	5.0E+15	3.1E+15	4.1E+15	3.1E+16
Freight loco 6x20 t, 120 km/h	60.3	149.6	3.9	979	0.10	0.07	0.00	0.45	967 401	123 351	916 497	1 691 515	5.3E+15	4.4E+15	4.2E+15	3.1E+16
Freight loco 6x30 t, 70 km/h	21.8	23.9	4.9	57	0.04	0.01	0.03	0.05	995 613	9 051	987 982	1 008 328	9.7E+15	2.8E+14	9.4E+15	1.0E+16
Passenger car 4x14 t, 160 km/h	81.5	154.3	4.0	900	0.16	0.13	0.00	0.70	756 207	139 106	681 773	1 511 770	3.6E+15	4.5E+15	2.3E+15	2.9E+16
Passenger Loco 4x19 t, 140 km/h	73.0	170.0	3.8	900	0.07	0.05	0.00	0.38	762 578	114 702	709 773	1 332 727	5.4E+15	5.0E+15	3.9E+15	3.1E+16
Passenger Loco 4x19 t, 175 km/h	60.2	130.0	4.4	764	0.10	0.06	0.00	0.34	770 689	114 165	719 119	1 332 727	5.5E+15	4.9E+15	4.0E+15	3.1E+16
Motor coach 4x16 t, 200 km/h**	40.8	68.6	3.6	557	0.14	0.11	0.02	0.58	877 980	93 333	747 933	1 242 539	5.9E+15	2.6E+15	3.7E+15	2.7E+16
Three-piece bogie 4x30 t, 60 km/h	87.1	21.6	55.8	102	0.25	0.15	0.13	0.42	867 067	3.6E-06	867 067	867 067	2.1E+16	3.1E+08	2.1E+16	2.1E+16
Three-piece bogie 4x6.5 t, 60 km/h	32.1	28.1	3.5	71	0.09	0.06	0.02	0.15	603 891	26 747	569 792	633 405	5.0E+14	2.3E+14	3.3E+14	8.5E+14
Freight wagon 2x22 t, 100 km/h	37.1	96.9	2.8	745	0.12	0.12	0.00	0.57	462 377	29 817	410 374	625 570	1.1E+16	3.2E+15	7.3E+15	3.2E+16
Freight wagon 2x6.5, 100 km/h	67.2	85.5	1.5	500	0.37	0.31	0.00	1.26	394 093	89 939	344 026	996 142	1.5E+15	2.8E+15	7.3E+14	2.2E+16
Motor coach 4x12 t, 140 km/h***	33.9	124.2	2.5	887	0.06	0.05	0.00	0.24	682 236	175 281	636 122	1 689 010	2.3E+15	4.5E+15	1.4E+15	2.8E+16
Motor coach 4x16 t, 200 km/h***	36.8	85.2	4.3	630	0.10	0.10	0.00	0.71	696 335	101 011	657 330	1 498 761	2.5E+15	2.6E+15	1.9E+15	2.4E+16
Motor coach 4x21 t, 200 km/h*,**	207.6	77.7	71.5	307	0.08	0.04	0.01	0.15	1 008 441	150 556	553 403	1 137 121	2.9E+16	8.3E+15	2.5E+15	3.2E+16
Motor coach 3x16.5 t, 160 km/h****	93.9	183.4	6.8	900	0.11	0.05	0.00	0.22	502 314	132 866	448 224	1 107 818	3.4E+15	5.6E+15	1.9E+15	3.0E+16
Motor coach 3x12.5 t, 200 km/h****	43.8	21.2	16.3	89	0.17	0.09	0.04	0.34	571 067	21 442	543 736	628 833	2.4E+15	2.6E+14	2.1E+15	3.1E+15
Y25 bogie 4x22 t, 100 km/h	58.3	158.3	4.5	1000	0.10	0.08	0.00	0.45	815 878	84 577	764 657	1 262 348	7.5E+15	4.7E+15	5.8E+15	3.3E+16

* High center of gravity, ** Stiff wheelset guidance, ***Flexible wheelset guidance, ***Jacob bogie and stiff wheelset guidance

Funded under the H2020 programme

Collaborative project H2020-MG-2015-2015 GA-636237

Needs Tailored Interoperable Railway – NeTIRail-INFRA

Deliverable D1.7
Incentives Final Report Annex 4
Marginal wear and tear costs in France using Bayesian analysis

Document ID: NeTIRail-WP1-D1.7v1.0-FINAL – ANNEX4

Due date of Deliverable: 30/09/2017

Actual submission date: 21/12/2017

Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Task leader for this deliverable: Professor Andrew Smith, Institute for Transport Studies, University of Leeds

Document status		
Revision	Date	Description
V0.1		First draft (authors Thijs Dekker, Phill Wheat, Andrew Smith)
V0.2	14/12/2017	
V1.0	21/12/2017	
Reviewed	YES	

Executive Summary

The purpose of this paper is to explore the extent to which a Bayesian approach to rail infrastructure cost modelling can help improve the robustness of such models. The Bayesian approach is expected to be particularly beneficial when relatively small sample sizes can be combined with prior information, obtained from previous cost modelling exercises in comparable contexts, on the parameters of interest. The Bayesian approach is also expected to be able to accommodate the use of flexible econometric specifications whilst maintaining the consistency of the model with economic theory, for instance, by imposing concavity of input prices.

We illustrate our work with an application to SNCF in France.

The overall contribution of this work to the state-of-the-art in empirical analysis of marginal wear and tear costs for rail infrastructure is:

- Demonstration of the approach and feasibility of Bayesian estimation in this context
- Demonstration of the benefit of Bayesian estimation: namely the ability to impose prior information on the implied elasticity relationship to exploit the findings from a large body of empirical studies in this area. This is potentially of great benefit when sample sizes are small
- Consideration of measures to determine whether the data under consideration is compatible with the prior information i.e. are the priors valid?
- Consideration of the benefits of this approach vis-à-vis a classical approach for different sample sizes, showing the importance of prior information for smaller sample sizes.

Approach

Bayesian analysis departs from classical statistical analysis (e.g. Ordinary Least Squares regression), since it assumes that model parameters are essentially random and that their distribution can be learned about through multiple studies which, when combined, produce the most appropriate estimate of the parameter given the sum of the information available at a given point in time.

This is very relevant to the issue of marginal cost for a number of reasons:

- There exists a large number of past studies undertaken across Europe. Further there exists established best practice synthesis such as Wheat et al (2009), which have produced recommendations for the values of the elasticity of cost with respect to traffic.
- Infrastructure managers, overseen by economic regulators, are required by EU legislation to set their track access charges based on an estimate of the marginal cost of running extra vehicles on the network. The legislation provides for different methods to be used, including econometric methods. A key question for infrastructure managers and regulators alike is first of all whether to undertake a new econometric study using their own country data or to rely entirely on past estimates. An econometric exercise is non-trivial in terms of data collection (particularly if the desire is to obtain a large enough sample) and also the skills required to undertake the exercise.
- Taking the above two points together, a further question – which is a key focus of this paper – is whether infrastructure managers / regulators could develop their own, bespoke econometric study, whilst utilizing the information contained in previous work in the form of

informed priors within a Bayesian framework. Such an approach could be particularly valuable in a situation where the bespoke dataset is small. Of course the necessary econometric skills would need to be available or bought in to implement such an approach.

In Bayesian analysis a parameter has a 'prior' distribution, which embodies what is known about the parameter before undertaking the study in question. In this paper, we use the Wheat et al (2009) recommendations for the range of appropriate values for the cost elasticity with respect to traffic as the prior information and then derive, through Bayesian estimation, the posterior distribution and thus an estimate using various splicing of a dataset from France. We consider three prior scenarios as shown in Table E.1.

Table E.1 Relationship between prior distribution and estimates from the posterior distribution

Prior Distribution Class	**Implementation in Section 5**	**Properties of estimates from Posterior Distribution**
Non-informative prior	Normal distribution with very large variance	Parameter estimates will be very similar to those from classical statistics e.g. OLS
Informative prior	Normal distribution N(0.2,0.01)	Parameter estimates are not forced to be within any bound however the prior will influence the estimates to the extent that the estimates will likely be closer to the prior mean relative to the OLS estimates. As the sample size increases, the influence of the prior diminishes, such that the estimates will approach the results from classical statistics
Informative prior – bounded parameter space	Uniform distribution. In section 5, we assume the traffic elasticity is Uniform[0.2,0.35]	Parameter estimates are forced within the bounds of the prior density ([0.2, 0.35] in section 5).

Findings

Our key conclusions are:

- We have been able to estimate Bayesian formulations of infrastructure cost functions for a dataset for France. Estimation results are comparable to those from classical approaches, however there is a clear influence of the prior information as intended.

- We have demonstrated that the influence of prior information for the posterior estimates is most influential when there is a limited sample size. This indicates that Bayesian analysis might be very beneficial when sample sizes are limited. Indeed, these techniques could be of great benefit where a country is considering developing an econometric study of marginal wear and tear costs for the first time (as the data requirements are less than trivial for a full classical study). Of course this assumes that the analyst and policy maker has confidence in the appropriateness of the prior information.

- We have outlined a measure of prior data conflict which highlights when the prior information is incompatible with the data in the sample. This is important for determining the appropriateness of the prior. Applying this criterion to our dataset reveals that bounded priors (fixed ranges of permissible values for an elasticity in our application) do lead to instances of prior data conflict. However given the nature of the generalisation framework in Wheat et al (2009) which involved judgement over a wide range of studies, unbounded priors are most appropriate.

This work represents a first exploration of the value of applying Bayesian techniques to the problem of marginal wear and tear cost estimation. Recommendations from this and subsequent work would hope to inform whether a infrastructure manager or economic regulator should rely only on their own available dataset to inform charges, or whether they should be explicitly drawing on past information, such as from the FP7 CATRIN project (Wheat et al, 2009), either as prior information in Bayesian econometric work or relying on past evidence only.

We consider that a natural extension of this work is to generalise the techniques to impose prior information on elasticities in second order cost functions e.g. Translog as these provide a more flexible descriptions of costs and represent the state-of-the-art. This involves assigning priors to both functions of model parameters and data, which is a subject for further research.

Table of contents

Abbreviations and acronyms

Abbreviation / Acronym	Description
MC	Marginal cost

1. Introduction

The purpose of this paper is to explore the extent to which a Bayesian approach to rail infrastructure cost modelling can help improve the robustness of such models. The Bayesian approach is expected to be particularly beneficial when relatively small sample sizes can be combined with prior information, obtained from previous cost modelling exercises in comparable contexts, on the parameters of interest. The Bayesian approach is also expected to be able to accommodate the use of flexible econometric specifications whilst maintaining the consistency of the model with economic theory, for instance, by imposing concavity of input prices.

We illustrate our work with an application to SNCF in France.

The overall contribution of this work to the state-of-the-art in empirical analysis of marginal wear and tear costs for rail infrastructure is:

- Demonstration of the approach and feasibility of Bayesian estimation in this context

- Demonstration of the benefit of Bayesian estimation: namely the ability to impose prior information on the implied elasticity relationship to exploit the findings from a large body of empirical studies in this area. This is potentially of great benefit when sample sizes are small

- Consideration of measures to determine whether the data under consideration is compatible with the prior information i.e. are the priors valid?

- Consideration of the benefits of this approach vis-à-vis a classical approach for different sample sizes, showing the importance of prior information for smaller sample sizes.

This work represents a first exploration of the value of applying Bayesian techniques to the problem of marginal wear and tear cost estimation. Recommendations from this and subsequent work would hope to inform whether an infrastructure manager or economic regulator should rely only on their own available dataset to form charges, or whether they should be explicitly drawing on past information, such as from the FP7 CATRIN project (Wheat et al, 2009), either as prior information in Bayesian econometric work or relying on past evidence only.

This is very relevant to the issue of marginal cost for a number of reasons:

- There exists a large number of past studies undertaken across Europe. Further there exists established best practice synthesis such as Wheat et al (2009), which have produced recommendations for the values of the elasticity of cost with respect to traffic.
- Infrastructure managers, overseen by economic regulators, are required by EU legislation to set their track access charges based on an estimate of the marginal cost of running extra vehicles on the network. The legislation provides for different methods to be used, including econometric methods. A key question for infrastructure managers and regulators alike is first of all whether to undertake a new econometric study using their own country data or to rely entirely on past estimates. An econometric exercise is non-trivial in terms of data collection (particularly if the desire is to obtain a large enough sample) and also the skills required to undertake the exercise.
- Taking the above two points together, a further question – which is a key focus of this paper – is whether infrastructure managers / regulators could develop their own, bespoke

econometric study, whilst utilizing the information contained in previous work in the form of informed priors within a Bayesian framework. Such an approach could be particularly valuable in a situation where the bespoke dataset is small. Of course the necessary econometric skills would need to be available or bought in to implement such an approach.

This paper is setup as follows. In section 2 the Bayesian approach is introduced in the context of the traditional Cobb-Douglas cost regression model. Section 3 and 4 discuss the prior and posterior densities respectively which are key components of the Bayesian approach. The empirical example is contained in section 5 and section 6 concludes.

2. A Bayesian approach to the Cobb-Douglas regression model

Equation (1) presents the standard Cobb-Douglas regression model where C_{it} denotes cost for track unit *i* in time *t*. Y_{it}represents track length and U_{it}traffic density. This functional form can be extended by including dummy variables capturing specific regional characteristics or by adding other explanatory variables, which are all captured by the vector $\mathrm{Z_{it}}$, to the model.[1] For notational convenience Z_{it} also contains the regression constant.

$$\ln(C_{it}) = \beta_{\mathrm{Y}} \cdot \ln(Y_{it}) + \beta_{\mathrm{U}} \cdot \ln(U_{it}) + \beta_{\mathrm{Z}} \cdot \mathrm{Z_{it}} + \epsilon_{it} \qquad (1)$$

Similar to the standard regression model, the Bayesian estimation framework requires assumptions regarding the distributional form of the error term ϵ_{it}. Like in any regression model, the error term is assumed to be identically and independently (i.i.d.) normally distributed with mean 0 and variance σ^2. Consequently, the likelihood function associated with equation (1) is *identical* in the maximum likelihood and Bayesian estimation framework.[2] The difference between the Bayesian and maximum likelihood framework arises in how the parameters β and σ^2 are interpreted. Maximum likelihood estimation assumes these parameters are fixed entities for which we obtain a sampling distribution, i.e. our parameter estimates vary depending on the sample used. Bayesian estimation instead assumes β and σ^2 are random entities for which all information (and uncertainty) is summarised by a set of probability distributions. A distinction is made between prior and posterior information. Prior information, or the prior density, describes the information we have about β and σ^2 before observing the data. After observing the data, we have additional information about these parameters through the likelihood function. Jointly the information contained by the prior density and the likelihood function is summarised by the posterior density. The posterior density arises by linking the prior density and the likelihood through Bayes rule, hence the term Bayesian estimation.

3. The prior density

Prior densities form an integral part of the Bayesian analytical framework and since they are by definition unrelated to the data, they are inherently subjective, or at least based on an information set independent of the data set for the study. The subjective nature of specifying prior densities has

[1] In the dataset Y and U are respectively operationalised by *is_lg_vo* and *tbc_voy*, *tbc_dhf* and *is_tdens*.

[2] The likelihood function is always written conditional on the parameters β and σ^2.

always been a criticism of the Bayesian approach and many classically trained econometricians feel uncomfortable with this aspect. Below, we will introduce some terminology regarding the prior before translating this into the context of our Cobb-Douglas model. From the outset it is important to highlight that the researcher can select *any* type of prior density and that when the sample size is large enough, the Bayesian estimation results will converge to the maximum likelihood results. The role of the prior will accordingly become irrelevant when using a sufficiently large sample size.

There are three cases when the prior becomes important. First, in small datasets there is relatively little information contained in the dataset and in those cases informed priors have a lasting effect on the posterior. Especially when estimating complex model specifications on small datasets the stability of the estimation procedure can be increased when the information contained in the prior aligns with the information in the data. Then the posterior is merely a refinement of the prior. However, and this is the second case, the prior information may not always be in line with the information contained by the data. This can result in reduced stability and biased outcomes of the estimation procedure. This is also known as prior-data conflict (Evans 2015). Finally, the prior can also be used to impose parameter restrictions on the model and thereby control the domain of the posterior. Namely, Bayes rule implies that when the prior assigns a zero probability to a particular value of the parameter, then that particular value should also have a zero probability in the posterior.

3.1 Non-informative, objective, reference and weakly informative priors:

Econometricians and statisticians have tried to minimise the influence the role of the prior on the posterior by using non-informative priors intending to make the analysis more objective by letting the data speak for themselves. The definition of what entails a non-informative prior is rather ambiguous and related terminology also refers to objective, reference and weakly informative priors respectively. Koop et al. (2007, p.80) define objective priors as those for "*which the contributions of the data are posterior dominant for the quantity of interest*". Key contributions on objective priors have been made by Jeffreys (1961). The so-called class of Jeffreys' priors lead to identical posterior inferences irrespective of the parameterisation of the model would. Jeffrey's priors have, however, been criticised by Kass and Wasserman (1996) arguing that, amongst other critiques, it merely evolved to selecting priors by convention rather than true objectivity. Bernardo (1979) introduced reference priors as an alternative to objective priors, noting that a reference prior maximises the missing information in an experiment. Typically, one would measure this information distance by the Kullback-Leibler distance between the prior and posterior density. In other words, reference priors ensure most weight (i.e. information) is given to (contained in) the likelihood. In many cases, the reference prior reduces to Jeffreys' prior for sufficiently large sample sizes. Finally, Evans and Jang (2011) refer to weakly informative priors as a set of priors that are purposely less informative than another set of priors including all the relevant information available to the analyst. The goal of weakly informative priors is to improve the stability and regularity of the obtained parameter estimates (Gelman et al. 2008). In other words, weakly informative priors fall in between reference priors and informative priors.

3.2 Prior-data conflict

The notion of weakly informative priors is closely related to the concept of prior-data conflict (Evans 2015). The intuition underlying Bayesian estimation is that the use of prior information generally results in more informative statistical inference. However, in certain cases the prior strongly contradicts the observed data such that the parameter values supported by the data all have (very) low prior probability. In these cases, there is a prior-data conflict. Evans and Jang (2011) argue that in such cases the priors should be adjusted, i.e. replaced by more weakly (conservative) informative priors reducing the probability of observing prior-data conflicts. This argument is, however, entirely driven by statistical arguments and goes against other potential reasons to adopt a particular prior; such as imposing parameter restrictions to ensure consistency of the estimated model and economic theory, which is one of the objectives of this study.

The problem of prior-data conflict is that information from the prior subtracts rather than adds to the analysis and thereby complicates the analysis. Unfortunately, there is no solution to this problem apart from adjusting the prior or testing the robustness and subsequently acknowledging the potential influence of using a particular prior. Nevertheless, it is a good thing to be aware of the potential influence of the prior on the posterior and the possible presence of prior-data conflict. In the context of this study, we wish to assess how influential the inclusion of prior information is on our posterior analysis, i.e. compare the informational content of the priors and position this relative to the informational content of the likelihood.

3.3 Measuring informational content and identifying prior-data conflict:

The Kullback-Leibler (KL) divergence criterion, defined in equation (2), is often used as a measure of the informational benefit in moving from the prior to the posterior. It is closely related to information theory by its close resemblance with the entropy measure:

$$D_{KL}\left(p(\beta,\sigma^2|Y),p(\beta,\sigma^2)\right)=\int\left(p(\beta,\sigma^2|Y)\ln\left(\frac{p(\beta,\sigma^2|Y)}{p(\beta,\sigma^2)}\right)\right)d\beta,\sigma^2 \tag{2}$$

Using Bayes rule, the D_{KL} reduces in (3) to the difference between the expected log-likelihood, where the posterior density acts as the weight in obtaining this expected value, and the log of the marginal likelihood. The prior influences D_{KL} through the posterior weights and the marginal likelihood.[3]

$$D_{KL}\left(p(\beta,\sigma^2|Y),p(\beta,\sigma^2)\right)=\int\left(p(\beta,\sigma^2|Y)\ln\left(p(Y|\beta,\sigma^2)\right)\right)d\beta,\sigma^2-\ln\left(p(Y)\right) \tag{3}$$

Bousquet (2008) makes use of a slightly altered version of the D_{KL} measure to test for prior-data conflicts. It takes the ratio of two D_{KL} measures. The denominator represents the D_{KL} from a model based on an uninformative prior. The numerator then calculates the D_{KL} using the same posterior as in the denominator whilst replacing the prior inside the ln() term in (2). If that ratio is larger than one a prior-data conflict is identified. Evans and Jang (2011), Nott et al. (2016) and Reimherr et al. (2014) define alternative measures and test statistics to assess the information divergence to assess prior informativeness and prior-data conflict. The applicability of these measures is not as developed and easy to apply as the Bousquet (2008) measure. All measures, more or less, identify a prior-data conflict when the D_{KL} divergence increases relative to the uninformative prior.

3.4 Specifying a prior for the Cobb-Douglas regression model:

Now that we are equipped with a better understanding of the notion and role of the prior, we can make this more explicit by defining a prior for the parameters in the Cobb-Douglas regression model in (1). That is, we need to make assumptions regarding the prior densities on β and σ^2.

[3] For our regression model, the marginal likelihood can be evaluated using Chib (1995) and Chib and Jeliazkov (2001) and the expected likelihood in itself is easy to evaluate at each draw from the Gibbs sampler.

Let's denote β as the vector containing all $\beta_Y, \beta_U, \beta_Z$ parameters and assume this vector to follow a multivariate normal density with prior mean μ_0 and prior covariance matrix Σ_0. The prior mean then controls the location of the prior density, e.g. on average what we think what the length elasticity β_Y should be. Such information would typically be based on previous studies. The prior variance, i.e. the diagonal elements of Σ_0 on β then controls for how precise our knowledge regarding β is. The off-diagonal elements of Σ_0 summarise how these parameters may relate to each other, but in most applications a diagonal prior covariance matrix is applied. The choice for a multivariate normal prior is driven by convenience. Namely, combining a normal prior with a normal likelihood function, like the linear regression model described by (1), results in an analytical (conditional) posterior from which it easy to take draws.[4] This will become clear later on when deriving the posterior densities. Alternative priors can be implemented, although this will significantly increase model run times and increases the potential for simulation noise.

The variance of the error term σ^2 needs to be strictly positive. This is recognised by introducing an inverse gamma prior.[5] The choice for the inverse gamma prior is again driven by convenience. Namely, the inverse gamma is a conjugate prior in this particular model setting as illustrated in the Appendix. The inverse gamma density is characterised by the shape parameter ω_0 and the scale parameter ν_0. Higher shape parameters shift the mode of the distribution closer to the origin and reduce the weight in the tail of the distribution, whereas the scale parameter controls the spread of the distribution.

4. The posterior density

Assuming C_{it}^* represents $\ln(C_{it})$ and X_{it} is the vector with all (transformed) explanatory variables, the posterior can be defined according to Bayes' rule by equation (4). Note that C^*represents the vector with all cost observations in the dataset and X the matrix of explanatory variables.

$$p(\beta,\sigma^2|C^*) = \frac{p(C^*|\beta,\sigma^2,X)p(\beta|\mu_0,\Sigma_0)p(\sigma^2|\omega_0,\nu_0)}{\int p(C^*|\beta,\sigma^2,X)p(\beta|\mu_0,\Sigma_0)p(\sigma^2|\omega_0,\nu_0)d\beta,d\sigma^2} = \frac{p(C^*|\beta,\sigma^2,X)p(\beta|\mu_0,\Sigma_0)p(\sigma^2|\omega_0,\nu_0)}{P(C^*|X)} \quad (4)$$

Since the denominator in (4) is independent of the model parameters of interest, i.e. these are integrated out, the denominator only acts as a normalising constant on the posterior density. Thereby it has no influence on the shape of the posterior and Bayesians therefore typically work with Equation (5). In words, (5) states that the posterior is proportional to the likelihood $p(C^*|\beta,\sigma^2,X)$ and the joint prior density $p(\beta|\mu_0,\Sigma_0)p(\sigma^2|\omega_0,\nu_0)$. Again, the likelihood function is identical to the one used in the maximum likelihood framework.

$$p(\beta,\sigma^2|C^*) \propto p(C^*|\beta,\sigma^2,X)p(\beta|\mu_0,\Sigma_0)p(\sigma^2|\omega_0,\nu_0) \quad (5)$$

4.1 Conditional posteriors and the Gibbs Sampler

The specified multivariate normal and inverse gamma priors allow us to work out two sets of analytical *conditional* posteriors, respectively $p(\beta|C^*,\sigma^2)$ and $p(\sigma^2|C^*,\beta)$ where the conditionality on X is suppressed for notational convenience. Appendix A analytically derives these two conditional

[4] This is also known as a conjugate prior.

[5] An alternative formulation is to work with the precision parameter $h = \frac{1}{\sigma^2}$ which follows a gamma prior.

posteriors highlighting they are respectively again a multivariate normal and inverse gamma density. These two conditional posterior densities form the basis for the Gibbs Sampler. The general idea behind the Gibbs Sampler (GS) is simple: break the joint posterior (Equation (5)) into conditional posteriors for which the analytical form of its density is known. Then sample sequentially and repeatedly from these conditionals. After a number of draws the joint sequence of conditional draws will converge to the desired joint posterior densities for all parameter. In addition, each individual sequence can be interpreted as the marginal posterior for a given parameter.

In our case the GS takes the following steps:

1. Set a starting value for σ^2
2. Conditional on σ^2 calculate
   $\Sigma_1 = \left(\Sigma_0^{-1} + (\frac{1}{\sigma^2})X'X\right)^{-1}$ – the posterior variance on β and
   $\mu_1 = \Sigma_1\left(\Sigma_0^{-1}\mu_0 + \left(\frac{1}{\sigma^2}\right)X'C^*\right)$ - the posterior mean on β
3. Take a draw for β from $\beta \sim MVN(\mu_1, \Sigma_1)$
4. Given the draw for β calculate:
   $\omega_1 = \omega_0 + \frac{n \cdot t}{2}$ – the posterior shape parameter for σ^2 and
   $\nu_1 = \nu_0 + \sum_i \sum_t \frac{(C_{it}^* - X_{it}\beta)^2}{2}$ the posterior scale parameter for σ^2
5. Take a draw for σ^2 from $\sigma^2 \sim IG(\omega_1, \nu_1)$
6. Repeat steps 2-5 a large number of times and, after a set of burn-in draws, store the retained draws for posterior analysis.

Some remarks can be made about the outcomes of steps 2. First, the posterior variance Σ_1 reduces to $\sigma^2(X'X)^{-1}$ when the prior variance on β approaches $+\infty$. The latter is also known as an improper uninformative prior. The posterior variance is then identical to the asymptotic OLS (and maximum likelihood) variance covariance matrix for β. Second, under the same improper uninformative prior the posterior mean μ_1 reduces to $(X'X)^{-1}X'C^*$, which again represents the OLS (and maximum likelihood) estimate. When prior information is introduced, the posterior variance goes down by definition compared to the OLS covariance matrix. This is a direct result of adding more information, which is equivalent to adding more data points to a dataset. Finally, the posterior mean is a weighted average of the prior mean and information contained in the data, where the weights are determined by the prior variance and the variance on the error term. An uninformative prior reduces the impact of the prior mean on the posterior mean. The final two points should be interpreted conditional on a given value for σ^2 which, however, changes across the iterations of the Gibbs Sampler and thereby adds a degree of noise to the joint posterior density.

For interpreting the outcome of step 4 we can refer to the MLE outcome for $\hat{\sigma}^2 = \frac{\sum_i \sum_t (C_{it}^* - X_{it}\beta)^2}{n \cdot t}$ and observe that the mean of the inverse gamma density is given by $\mu_{IG} = \frac{\nu_0 + \sum_i \sum_t \frac{(C_{it}^* - X_{it}\beta)^2}{2}}{\omega_0 + \frac{n \cdot t}{2} - 1}$, which for $\omega_0 = 1$ and $\nu_0 = 0$ would reduce to $\hat{\sigma}^2$. Indeed, a lower prior scale parameter increases the variance of the prior and will be extremely uninformative for being ν_0 close to zero. However, ν_0 needs to take some positive value. Also for the prior inverse gamma density to have a mean and a variance $\omega_0 > 2$ is required. Hence, the addition of prior information induces some differences between the Bayesian and classical estimate for σ^2, but this difference is rapidly reducing in sample size. More details on this specification can be found in Koop (2003) Chapter 4.2.

4.2 Summary of relationship between prior and posterior density

Sections 3 and 4 have discussed the prior and posterior densities respectively. Table 1 summaries the key results of how the choice of prior influences the posterior density and thus the parameter estimates. There are three cases: a non-informative prior, an informative prior which spans all the possible parameter space (e.g. a normal distribution) and an informative prior which restricts the parameter space i.e. a uniform distribution. All results assume that the same class of priors are applied to all parameters.

Table 1 Relationship between prior distribution and estimates from the posterior distribution

Prior Distribution Class	Implementation in Section 5	Properties of estimates from Posterior Distribution
Non-informative prior	Normal distribution with very large variance	Parameter estimates will be very similar to those from classical statistics e.g. OLS
Informative prior	Normal distribution N(0.2,0.01)	Parameter estimates are not forced to be within any bound however the prior will influence the estimates to the extent that the estimates will likely be closer to the prior mean relative to the OLS estimates. As the sample size increases, the influence of the prior diminishes, such that the estimates will approach the results from classical statistics
Informative prior – bounded parameter space	Uniform distribution. In section 5, we assume the traffic elasticity is Uniform[0.2,0.35]	Parameter estimates are forced within the bounds of the prior density ([0.2, 0.35] in section 5).

5. Empirical Application

In the preceding sections, we have specified the traditional Cobb-Douglas regression model in a Bayesian setting. We have discussed the role of the prior and specified a specific distributional form. Based on this specific distributional form, we were able to derive the corresponding posterior density.[6] We will now move on to an empirical illustration of the concept by contrasting the results from a traditional OLS model with that of the corresponding Bayesian model. We will do so by using three different priors. The priors we will be using are respectively an uninformative prior for which we expect the results to confirm those of the OLS model; two sets of informative priors which vary in shape and location.

In terms of using prior information, our main explanatory variable of interest is the size variable Total Traffic. Since the dependent and explanatory variable are both specified in terms of the natural logarithm, the corresponding coefficient represents the elasticity. Based on previous research in the

[6] It should be kept in mind that the posterior density will change when the underlying prior density changes.

FP7 CATRIN project (Wheat et al, 2009), we expect this elasticity to fall within the range between 0.2 and 0.35. The priors on the remaining explanatory variables will remain uninformative.

5.1 Data used

We apply Bayesian techniques a dataset of infrastructure maintenance for track assets in the French railway network. The dataset comprises of 1370 track sections in France for a single year, 2013. The dataset has previously been analysed by Walker et al (2017). The Walker et al analysis used classical techniques but considered more flexible functional forms (Translog and Box-Cox). This analysis is not a substitute for that work, since we adopt a similar model specification. It does however illustrate the potential advantages of adopting a Bayesian approach in terms of key results even if we acknowledge the scope for further work (see conclusions).

The final sample for estimation smaller than the full 1370 Due to missing data and identification of outliers in the data we adopt the same sample as Walker et al (2017) with 1128 observations in total for the track maintenance model.

The dependent variable is track maintenance cost per track km. The explanatory variables comprise a traffic variable, namely gross tonne-km per track-km and a set of control variables which characterise the inherent quality and performance of the infrastructure:

- Average age of rail (years)
- Average maximum allowed speed
- Number of switches and crossings per track km
- Average number of sleepers per track km
- Proportion of track that is in a curve
- Proportion of continuous welded rail
- Dummy for high-speed traffic tracks
- Regional dummies

5.2 Results

Table 2 presents the results for the OLS and the Bayesian models referred to above. Notably, at 0.246 the estimated OLS elasticity for Total Traffic falls well within the range where we would expect the elasticity to be. This result is lower than that found in Walker et al (2017), who used Translog models and Box Cox models which are more flexible. Again it should be noted that this is a Cobb-Douglas model for the purpose of illustration of the Bayesian approach. We consider opportunities for further work to extend the techniques to second order functional forms in the conclusion section.

The other coefficients in the model are also of the expected sign. As a reminder, the model specification does not change across the three Bayesian models, only the assumptions regarding the prior distribution.

The first Bayesian model imposes an uninformative but proper prior, i.e. having a high but finite prior variance. The prior covariance matrix for β is therefore set to $\Sigma_0 = 1000 \cdot I$ where *I* represents the identity matrix and the prior mean is set to $\mu_0 = 0$.[7] Similarly, setting $\omega_0 = 2.01$ and $\nu_0 = 1.01$ translates into a prior mean of 1 and variance of 100 for the inverse gamma prior on σ^2. Given these

[7] With such a high prior variance the location of the mean is largely irrelevant.

specifications, 20,000 iterations of the Gibbs Sampler are then conducted of which only the second 10,000 draws are eventually used for posterior analysis. The initial set of 10.000 draws is discarded as burn-in draws. Convergence of the Gibbs Sampler is assessed graphically and using the Geweke (1992) convergence test.[8] As can be observed from Table 2, the prior was indeed uninformative and differences from the OLS results only arise at the third digit. Moreover, Figure 1 illustrates that the Gibbs Sampler has no convergence issues (i.e. is stable) and suffers little from simulation noise (i.e. does not linger in specific areas of the posterior for a longer period of time. This is also confirmed for the other model parameters. In terms of model fit statistics, there is a large discrepancy between the expected log-likelihood and the marginal likelihood (also reported in logs). This is a direct result of working with a large number of parameters for which uninformative priors are specified. As to be expected, the Kullback-Leibler measure indicates there is a large informational distance between the prior and the posterior, i.e. a substantial amount of information is included in the data allowing to refine our knowledge regarding the model parameters as summarised by the posterior.

It is not possible to conduct conventional significance test based on these draws, but it is an easy exercise to identify the percentage of draws which are on the other side of zero thereby reflecting a one-sided test on the sign of the parameter. More informative, however, are so-called highest posterior density intervals (HPDI) denoting the smallest region in the posterior containing 95% of the density. This is the Bayesian equivalent of a confidence interval. In our example, we can state with 95% confidence that the elasticity falls between 0.188 and 0.286.

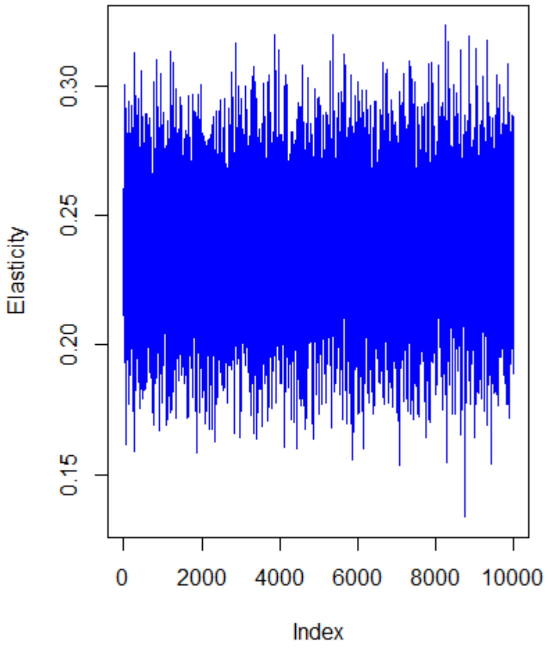


Figure 1: Draws for the elasticity on Total Traffic using an uninformative prior

[8] The Geweke (1992) test identifies whether the mean of the first half of the iterations used for analysis is significantly different from the second half and thereby assesses the stability of the posterior density.

The second Bayesian model introduces an informed prior on the elasticity for Total Traffic such that it's prior mean is assumed to be 0.3 and its prior standard deviation 0.1. The same prior assumptions and estimation settings as applied in the first Bayesian model apply to all other parameters. Due to the large number of observations in the current dataset and the correspondence of the OLS result with this informative prior, the results are not too different from the first Bayesian model. The posterior mean for the elasticity on Total Traffic displays a slight increase from 0.237 to 0.241 and the posterior standard deviation decreases slightly by the inclusion of additional (and most important accurate) prior information. The 95% confidence interval now ranges between 0.193 and 0.289 suggesting that primarily the lower bound has moved upwards compared to the non-informative case.

In the case of this model, we see an improvement in the marginal likelihood compared to the first Bayesian model. Taking the difference between these two marginal likelihoods represents the log of what is known as a Bayes Factor (Kass and Raftery 1995) a key tool for model comparison of nested and non-nested models. This improvement in fit is, however, not confirmed by the expected log-likelihood. That is, on average the model is not able to better explain the data, but experiences a lower penalty in the marginal likelihood penalty due to including one less uninformative prior. This supports the observation that for this particular dataset the data dominates the results and that the prior thereby has limited influence. Finally, the Bousquet (2008) measure confirms that we are not experiencing a prior data conflict. Also it does not come as a surprise that the measure is close to one since we've only included prior information on a single model parameter.

The third Bayesian model is somewhat more challenging in terms of its estimation. Here, we move away from the normal conjugate prior on Total Traffic and impose a uniform prior ranging between 0.2 and 0.35 instead. By restricting the domain of the prior, we automatically restrict the domain of the posterior to that same range. In effect, we thus declare a very strong prior on which values for the elasticity are allowed. Moreover, using a uniform prior breaks the conjugacy property of the prior on this parameter such that the conditional posterior is no longer of a convenient analytical form and a Metropolis-Hastings algorithm (MH) needs to be used to obtain draws from the conditional posterior density of interest (Chib and Greenberg 1995; Train 2009). An additional step is added to the Gibbs Sampler and a significantly larger number of draws (10.000 burn-in draws and 70.000 retained draws) needs to be applied to account for the fact that the MH procedure is far less efficient due to the presence of serial correlation and since not all candidate draws will be accepted. Both of these are clearly reflected in Figure 2. The draws look, however, stable enough and are moving around sufficiently random over the parameter space that there is no need for concern.

Again the results confirm that the choice of prior has limited impact on the model results. It is not surprising that by restricting the parameter space the posterior standard deviation has decreased somewhat and that the mean has moved up slightly. Also, the 95% confidence interval, which ranges between 0.200 and 0.280 shows a significant overlap with the preceding confidence intervals but does indicate the lower bound was likely to be too restrictive (since the lower boundary is at the boundary imposed by the prior). The existence of a prior data conflict in this case is furthermore reflected by the decrease in the marginal likelihood compared to the second case, but it's most evident since 6.9% of the draws from our uninformative prior are outside the range of the uniform prior. Accordingly, the Bousquet (2008) measure cannot be calculated for this specific model but would by definition be >1 since the data suggests part of the posterior domain based on an uninformative prior falls outside of the domain based on a (very restrictive) informative prior. These results do not provide indications for concern due to the limited number of draws violating the domain of the uniform prior.

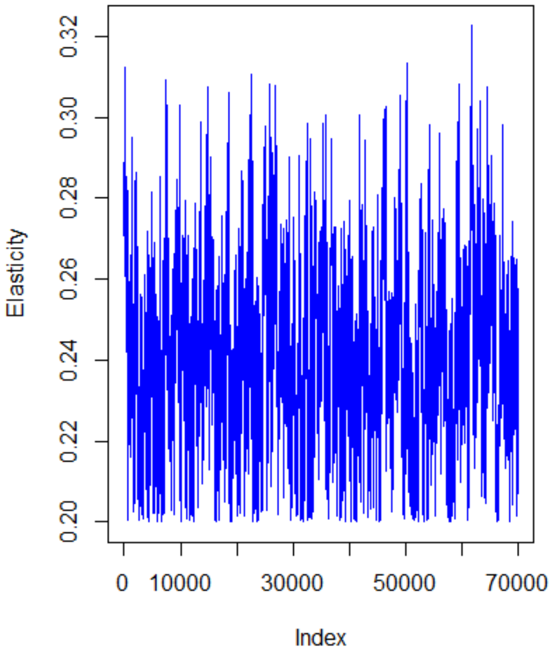


Figure 2: MH-draws for the elasticity on Total Traffic using a uniform prior

Table 2 Results of various estimation approaches using the full sample

	OLS		Uninformed prior		Informed prior		Uniform prior	
	Est	Std. Er.	Post mean	Post. Std.	Post mean	Post. Std.	Post mean	Post. Std.
constant	-1.442	0.741	-1.447	0.739	-1.462	0.739	-1.456	0.740
Total Traffic (ln)	0.237	0.025	0.237	0.025	0.241	0.024	0.242	0.022
Percentage of track curved	0.636	0.145	0.634	0.144	0.633	0.144	0.636	0.144
Percentage of track welded	-0.895	0.134	-0.894	0.133	-0.899	0.133	-0.900	0.133
LGV	-0.852	0.181	-0.852	0.182	-0.846	0.181	-0.847	0.179
Zero_pass_traffic	0.538	0.123	0.540	0.122	0.536	0.122	0.534	0.123
Zero_freight_traddic	0.920	0.326	0.923	0.326	0.920	0.326	0.916	0.326
Zero_S&C_per_km	0.060	0.127	0.061	0.127	0.060	0.127	0.058	0.127
Avg_dens_sleeper	0.001	0.000	0.001	0.000	0.001	0.000	0.001	0.000
Max_speed	0.005	0.001	0.005	0.001	0.005	0.001	0.005	0.001
Avg_age	0.008	0.002	0.008	0.002	0.008	0.002	0.008	0.002
S&C_per_km	0.137	0.033	0.137	0.033	0.136	0.033	0.136	0.033
o_reg_al	-0.058	0.168	-0.059	0.167	-0.059	0.167	-0.058	0.168
o_reg_aq	-0.278	0.161	-0.278	0.161	-0.276	0.161	-0.276	0.161
o_reg_au	-0.386	0.181	-0.388	0.181	-0.385	0.181	-0.383	0.181
o_reg_bn	-0.599	0.259	-0.603	0.258	-0.598	0.258	-0.593	0.259
o_reg_bo	-0.377	0.156	-0.380	0.157	-0.378	0.157	-0.375	0.155
o_reg_br	-0.413	0.221	-0.411	0.224	-0.410	0.224	-0.412	0.221
o_reg_ce	-0.149	0.149	-0.150	0.150	-0.149	0.150	-0.147	0.149
o_reg_ca	-0.296	0.157	-0.297	0.156	-0.296	0.156	-0.295	0.158

o_reg_fc	-0.082	0.233	-0.082	0.235	-0.079	0.235	-0.079	0.233
o_reg_hn	-0.112	0.181	-0.114	0.182	-0.111	0.182	-0.110	0.181
o_reg_id	0.007	0.121	0.005	0.120	0.003	0.120	0.004	0.120
o_reg_lr	0.288	0.185	0.285	0.186	0.284	0.186	0.287	0.185
o_reg_li	0.197	0.222	0.197	0.223	0.200	0.223	0.201	0.222
o_reg_lo	-0.479	0.146	-0.481	0.146	-0.481	0.146	-0.480	0.146
o_reg_mp	-0.340	0.172	-0.343	0.172	-0.339	0.172	-0.336	0.172
o_reg_np	-0.314	0.133	-0.316	0.134	-0.316	0.134	-0.314	0.133
o_reg_pl	-0.239	0.174	-0.240	0.174	-0.239	0.174	-0.238	0.175
o_reg_pi	-0.264	0.153	-0.269	0.153	-0.268	0.153	-0.264	0.153
o_reg_pc	-0.313	0.193	-0.314	0.190	-0.312	0.190	-0.312	0.193
o_reg_pa	0.013	0.165	0.012	0.165	0.012	0.165	0.012	0.166
sigma2	0.855		0.855	0.037	0.855	0.037	0.855	0.037
R^2 – Marginal likelihood (ln)	0.93		-1691.52		-1685.978		-1689.146	
Expected log-likelihood			-1512.614		-1512.592		-1512.474	
D_{KL}			178.906		173.386		176.6717	
Bousquet (2008) measure					0.96	No prior-data conflict	NA	Prior data conflict

5.3 Reduced sample size

The analysis conducted so far illustrated that the size of the dataset and the correspondence between the information in the dataset and our prior information does not really built a case for using Bayesian analysis. To this end, we decided to repeat the above analysis using a random sample of 100 observations from the original dataset which consists of 1128 observations. Table 3 presents the results.

As can be expected with using a smaller sample, not all the explanatory variables can be maintained in the model. For example, we no longer have observations from a specific region (*o_reg_pc*) and can therefore no longer estimate the respective coefficients. The regression model provides an error report and ignores the respective parameters when estimating the model. In Bayesian models, identification issues are not always apparent and the coefficients can still be estimated. Since the data provides no information on the role of the respective explanatory variables, the corresponding standard errors are very high, making the above-mentioned identification issues easily identifiable.

The use of a smaller sample increases the size of the standard error on all parameters including the elasticity on Total Traffic. The corresponding 95% confidence interval for the uninformative Bayesian model now stretches between 0.066 and 0.420 and the posterior mean still falls well within the expected range. The OLS model and the uninformative Bayesian model still show highly comparable results.

When moving to the informative Bayesian model using the normal prior, we see a noticeable increase in the posterior mean, i.e. moving closer to the prior mean of 0.3, and a reduction in the posterior standard deviation. The 95% confidence interval is much tighter, especially on the lower end, and ranges between 0.137 and 0.397. This clearly highlights that the prior has a lasting impact on the posterior. If we would have a lot of confidence in this prior this would translate in making better informed decisions as opposed to working solely upon the information contained in this small sample. Similar to the models based on the full sample size, the informative normal prior improves the marginal likelihood, but in this case also leads to an improvement in the expected log-likelihood. Hence, the benefits do not only arise in reducing the penalty for using uninformative priors but also in being (on average) better able to explain the data. As expected, we do not observe an indication of prior-data conflict based on this informative prior and it remains close to one due to only including prior information on a single parameter.

Moving to the model based on the uniform prior, we see that the posterior mean of the elasticity on Total Traffic is slightly higher than with the informed prior based on the normal density. We also observe a rather low posterior standard deviation which results in a tighter 95% confidence interval ranging between 0.200 and 0.339. Again, this is an indication that particularly the lower bound on the prior may have been set slightly too low. Not surprisingly, in this case there is even clearer evidence of a data conflict. No less than 42.9% of the draws from the posterior of the uninformative prior are not within the range of the (restrictive) uniform prior. The restrictiveness of this prior is highlighted by a slightly worse marginal likelihood compared to the model based on the normal informed prior. The learning effect, i.e. the informational distance between the prior and the posterior is also larger than for the second Bayesian model.

In all, this modelling exercise has shown that the use of prior information can help in making more accurate inference on the parameters of interest when faced with a small sample size. Moreover, the use of an unbounded informative normal prior appears to be less restrictive than the use of a uniform prior. The latter is more likely to lead to prior data conflicts leaving the analyst with the decision to define how much faith to put in such a restrictive prior. In the case of this application, it difficult to support a bounded prior given the underlying studies in Wheat et al (2009) did include instances of results outside of the [0.2, 0.35] range.

Table 3 Results from various estimation methods using a sample size of 100

	OLS		Vague prior		Informed prior		Uniform prior	
	Estimate	Std. Er.	Post mean	Post. Std.	Post mean	Post. Std.	Post mean	Post. Std.
constant	-2.131	2.032	-1.297	22.454	-1.370	22.453	-1.051	22.367
Total Traffic (ln)	0.246	0.090	0.245	0.090	0.269	0.066	0.273	0.040
Percentage of track curved	1.589	0.458	1.582	0.467	1.630	0.450	1.640	0.432
Percentage of track welded	-0.586	0.451	-0.577	0.457	-0.607	0.449	-0.622	0.439
LGV	-3.120	0.725	-3.118	0.732	-3.130	0.730	-3.127	0.723
Zero_pass_traffic	0.523	0.419	0.524	0.420	0.488	0.410	0.481	0.403
Zero_freight_traddic	NA	NA	-0.823	22.460	-0.897	22.460	-1.236	22.367
Zero_S&C_per_km	0.227	0.421	0.224	0.426	0.226	0.424	0.227	0.420
Avg_dens_sleeper	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
Max_speed	0.008	0.003	0.008	0.003	0.007	0.003	0.007	0.003
Avg_age	0.017	0.007	0.017	0.007	0.017	0.007	0.017	0.007
S&C_per_km	0.120	0.128	0.122	0.130	0.117	0.129	0.115	0.127
o_reg_al	0.092	0.473	0.091	0.480	0.094	0.479	0.098	0.474
o_reg_aq	-0.320	0.411	-0.317	0.416	-0.306	0.413	-0.311	0.409
o_reg_au	-1.536	0.694	-1.538	0.700	-1.524	0.697	-1.523	0.696
o_reg_bn	-0.555	0.493	-0.554	0.499	-0.527	0.493	-0.528	0.486
o_reg_bo	-0.537	0.428	-0.534	0.430	-0.532	0.429	-0.536	0.429
o_reg_br	2.278	0.992	2.282	1.005	2.242	0.998	2.230	0.986
o_reg_ce	-0.459	0.422	-0.459	0.421	-0.435	0.416	-0.432	0.414
o_reg_ca	-0.402	0.375	-0.400	0.379	-0.394	0.378	-0.397	0.376

o_reg_fc	0.727	0.864	0.724	0.875	0.728	0.873	0.732	0.868
o_reg_hn	-1.341	0.498	-1.349	0.508	-1.344	0.506	-1.335	0.502
o_reg_id	0.376	0.434	0.371	0.435	0.347	0.430	0.348	0.429
o_reg_lr	1.327	0.566	1.324	0.576	1.341	0.573	1.344	0.565
o_reg_li	0.679	0.664	0.683	0.668	0.730	0.656	0.727	0.649
o_reg_lo	-0.333	0.359	-0.341	0.362	-0.341	0.361	-0.334	0.361
o_reg_mp	-0.644	0.563	-0.644	0.567	-0.617	0.561	-0.616	0.559
o_reg_np	-0.301	0.571	-0.306	0.573	-0.261	0.561	-0.255	0.553
o_reg_pl	0.039	0.908	0.051	0.914	0.046	0.911	0.033	0.910
o_reg_pi	-0.460	0.420	-0.469	0.424	-0.460	0.422	-0.449	0.420
o_reg_pc	NA	NA	0.066	31.375	0.066	31.375	-0.141	31.821
o_reg_pa	0.480	0.477	0.481	0.475	0.516	0.466	0.514	0.464
sigma2			0.647	0.109	0.643	0.108	0.641	0.107
R^2 – Marginal likelihood (ln)	0.96		-254.779		-249.410		-252.450	
Expected log-likelihood			-119.847		-119.580		-119.303	
D_{KL}			134.932		129.826		133.150	
Bousquet (2008) measure					0.956	No prior data conflict		Prior data conflict

6. Conclusions and Further work

In this paper, we have demonstrated that Bayesian analysis can be implemented in the context of marginal wear and tear cost estimation in rail. The Bayesian approach provides promising results to make use of respectively prior information and theoretical expectations. It goes without saying that the Bayesian approach is able to combine prior information on parameters with theoretical expectations in a single model.

We have departed from standard implementations of Bayesian econometrics by utilising informative priors based on the results of past research and a subsequent research synthesis exercise reported in Wheat et al (2009). We have examined the importance of sample size as a way of determining how influential the prior information is on the resulting estimates under different circumstances. Finally we have considered the influence of a bounded prior as opposed to an unbounded prior and drawn on measures of the extent of prior data conflict to show the appropriateness of priors.

Our key conclusions are:

- We have been able to estimate Bayesian formulations of infrastructure cost functions for a dataset for France. Estimation results are comparable to those from classical approaches, however there is a clear influence of the prior information as intended.

- Demonstrated that the influence of prior information for the posterior estimates is most influential when there is a limited sample size. This indicates that Bayesian analysis might be very beneficial when sample sizes are limited. Indeed, these techniques could be of great benefit where a country is considering developing an econometric study of marginal wear and tear costs for the first time (as the data requirements are less than trivial for a full classical study). Of course this assumes that the analyst and policy maker has confidence in the appropriateness of the prior information.

- We have outlined a measure of prior data conflict which highlights when the prior information is incompatible with the data in the sample. This is important for determining the appropriateness of the prior. Applying this criterion to our dataset reveals that bounded priors (fixed ranges of permissible values for an elasticity in our application) do lead to instances of prior data conflict. However, given the nature of the generalisation framework in Wheat et al (2009) which involved judgement over a wide range of studies, unbounded priors are most appropriate.

6.1 Further research: Imposing prior information in flexible functional form

This application has assumed a simple first order functional form. This was because theoretical expectations are formulated in terms of the elasticity. The Cobb-Douglas functional form used in the empirical application directly estimates these elasticities. Imposing theoretical restrictions is therefore an easy task in the used model specifications. As discussed in the results section, this is simplification relative to the classical analysis in Walker et al (2017) who used second order functional forms. The elasticities in these models are functions of both multiple parameters and the data and so this adds complexity to the problem of imposing prior information on these relationships.

The Bayesian framework enables the researcher to implement parameter restrictions through the prior. For example, the inverse gamma density was selected for the variance of the error term to reflect that it needs to be positive. The uniform density used in the third Bayesian model implemented specific restrictions on the estimated model, but can easily be altered to reflect strictly positive values for the respective variable. The only downside of implementing such parameter restrictions is that most often the conjugate nature of the model specification is broken and the researcher needs to revert to the MH-algorithm. Our third Bayesian model acts as a case in point of braking the conjugacy of the model. Working with the MH-algorithm is not a problem as such, but it significantly increases the computational costs (and tuning efforts) of the model.

Implementing theoretical restrictions in alterative and more flexible functional forms where the elasticity (or other object of interest) is a function of multiple model parameters (and often explanatory variables) is slightly more complicated. This has received significant attention in the literature and Bayesian econometrics has proven itself as a useful framework in this context (e.g. Chalfant and Wallace 1992, Terrell 1996, Griffiths et al. 2000, O'Donnell and Coelli 2005, Wolff 2016).[9] All the referred papers impose a constraint on the prior and thereby implicitly on the posterior. They do so by evaluating whether for a possible set of parameter values in the domain of the prior in combination with the observed values for the corresponding explanatory variables the theoretical expectations are satisfied. If for a given set of parameter values the theoretical conditions are violated, the prior density (conditional on the value for the exogenous and observable explanatory variables) will be set to zero. That is, the multivariate prior density is effectively truncated in specific regions.

The practical implementation is less complicated. Since the analyst makes use of the MH-algorithm, i.e. (s)he needs to take a candidate draw and decide whether to accept or reject it. (S)he can simply validate whether the theoretical conditions are satisfied (or not) for that specific draw and then accept or reject the candidate draw and proceed accordingly (Terrell 1996). This requires a small extra step to the MH-algorithm. If the theoretical conditions are violated over a large area, the MH-algorithm may become rather inefficient as it is harder to find an acceptable candidate draw. Depending on the functional form and imposed prior densities the MH-algorithm can potentially be made more efficient (e.g. O'Donnell and Coelli 2005).

7. References

Bernardo, J. 1979. Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society. Series B,* 41(2), 113-147.

Bousquet, N. 2008. Diagnostics of prior-data agreement in applied Bayesian analysis. *Journal of Applied Statistics*, 35(9), 1011-1029.

Chalfant, J. and Wallace, N. 1992. Bayesian Analysis and regularity conditions on flexible functional forms: application to the US motor carrier industry. In: Griffiths et al. : *Readings in econometric theory and practice: a volume in honour of George Judge*, North-Holland, Amsterdam.

Chib, S. 1995. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90(432), 1313-1321.

Chib, S. and Greenberg, E. 1995. Understanding the Metropolis-Hastings Algorithm, *The American Statistician*, 49(4), 327-335.

[9] A key reference in the context of the classical estimation framework is the work by Diewert and Wales (1987).

Chib, S. and Jeliazkov, I. 2001. Marginal likelihood from the Metropolis–Hastings output. *Journal of the American Statistical Association,* 96(453), 270-281.

Diewert, W. and Wales, T. (1987) Flexible functional forms and global curvature conditions. *Econometrica*, 55(1), 4-68.

Evans, M. 2015. Measuring statistical evidence using relative belief. *Monographs on statistics and applied probability*; 144. CRC Press.

Evans, M. and Jang, G. 2011. Weak informativity and the information in one prior relative to another. *Statistical Science*, 26(3), 423-439.

Gelman, A. Jakulin, A. Pittau, M. and Su, Y 2008. A weakly informative default prior distribution for logistic and other regression models. *Annual Applied Statistics*, 2, 1360-1383.

Geweke, J. 1992. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics* 4, Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M. (eds.), 169-193. Oxford University Press.

Griffiths, W. O'Donnell, C. and Cruz, A. 2000. Imposing regularity conditions on a system of cost and factor share equations. *Australian Journal of Agricultural and Resource Economics*, 44(1), 107-127.

Jeffreys, H. 1961. *Theory of probability*. Oxford University Press.

Kass, R. and Raftery, A. 1995. Bayes Factors, *Journal of the American Statistical Association*, 90(430), 7739-795.

Kass, R. and Wasserman L. 1996. The selection of prior distributions by formal rules. *Journal of the American Statistical Society*, 91(453), 1343-1370.

Koop, G. 2003. *Bayesian Econometrics*. Wiley & Sons Ltd.

Koop, G. Tobias, J. and Poirier, D. 2007. Bayesian econometric methods. *Econometric Exercises*; 7. Cambridge University Press.

O'Donnell, C. and Coelli T. 2005. A Bayesian approach to imposing curvature on distance functions, *Journal of Econometrics,* 126, 493-523.

Nott, D. Xueou, W. Evans, M. and Englert, B. 2016. Checking for prior-data conflict using prior to posterior divergences. Version 28th of November 2016. **arXiv:1611.00113**

O'Donnell, C. and Coelli, T. 2005. A Bayesian approach to imposing curvature on distance functions. *Journal of Econometrics*, 126, 493-523.

Reimherr, M. Meng, X. and Nicolae, D. 2014. Being an informed Bayesian: Assessing prior informativeness and prior likelihood conflict. Version 23rd of June 2014. **arXiv:1406.5958**

Terrell, D. 1996. Incorporating monotonicity and concavity conditions in flexible functional forms. *Journal of Applied Econometrics*, 11, 179-194.

Train, K. 2009. *Discrete choice models with simulation*. Cambridge University Press.

Walker P, Smith, A. S.J., Wheat, P.E. and Marti, M. (2017). Modelling railway infrastructure maintenance cost in France: Overview of estimates, Report for SNCF Réseau.

Wheat, P., A.S.J. Smith, and C. Nash (2009): ‘CATRIN (Cost Allocation of TRansport INfrastructure cost)’, Deliverable 8 – Rail Cost Allocation for Europe, VTI, Stockholm.

Funded under the H2020 programme

Collaborative project H2020-MG-2015-2015 GA-636237

Needs Tailored Interoperable Railway – NeTIRail-INFRA

Deliverable D1.7
Incentives Final Report - Annex 5 – Marginal wear and tear costs in Sweden using a 16 year panel

Document ID: NeTIRail-WP1-D1.7v1.0-FINAL – ANNEX5

Due date of Deliverable: 30/09/2017

Actual submission date: 21/12/2017

Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Task leader for this deliverable: Professor Andrew Smith, Institute for Transport Studies, University of Leeds

Document status		
Revision	Date	Description
V0.1		First draft (authors Kristofer Odolinski and Phillip Wheat)
V0.2	14/12/2017	Review by USFD and ALU-FR
V1.0	21/12/2017	Final version
Reviewed	YES	

Executive Summary

The purpose of this paper is to provide empirical evidence on the wear and tear costs of rail infrastructure in Sweden, using a 16 year panel dataset. To do this we consider a dynamic panel data specification which allows for interdependence between maintenance and renewals, as well as their intertemporal effects. The estimates can be used to calculate the marginal cost for traffic, which has become an important part of the track access charges that were introduced after the vertical separation between train operations and infrastructure management in Europe as of the 1990s. Given that there are dynamic effects between different activities in infrastructure provision, the marginal cost estimates that takes these effects into account will be closer to the actual cost of running one extra unit of traffic on the railway, compared to the cost estimates based on static models for maintenance (see for example Wheat et al. 2009) and renewals (see for example Andersson et al. 2012 and Andersson and Björklund 2012).

Methodology

In this paper we analyze the dynamics between rail infrastructure renewals and maintenance in Sweden, using a panel vector autoregressive model. The dynamics in maintenance and renewals implies that an infrastructure manager (IM) needs to strike a balance within and between these activities for a certain traffic level. A sudden increase in traffic may thus require an adjustment of these costs. This implies that the cost impact from traffic needs to be studied in a dynamic context. The model estimation also comprises intertemporal effects for each of these activities.

We consider a panel VAR(p) model, where p denotes the lag length used in the model.[1] We have two endogenous variables: renewal costs (R_{it}) and maintenance costs (M_{it}), where $i = 1,2 \dots, N$ contract areas and $t = 1,2,\dots,T$ years. $\alpha_{1,i}$ and $\alpha_{2,i}$ are the unobserved individual-specific effects for the renewal and maintenance equations respectively, while $u_{1,it}$ and $u_{2,it}$ are their respective residuals, where $\left(u'_{1,it}, u_{2,it}\right) = \boldsymbol{u}_{it} \sim iid(0, \sum)$. $\sum$ is the covariance matrix of the errors. We also include a vector of exogenous variables $\boldsymbol{X}_{it}$ with parameters $\boldsymbol{\beta}_{11}$ and $\boldsymbol{\beta}_{21}$ for the maintenance and renewal equations respectively.

$$R_{it} = \alpha_{1,i} + \delta_{11} R_{it-1} + \theta_{11} M_{it-1} + \boldsymbol{\beta}_{11} \boldsymbol{X}_{it} + u_{1,it}$$

$$M_{it} = \alpha_{2,i} + \delta_{21} R_{it-1} + \theta_{21} M_{it-1} + \boldsymbol{\beta}_{21} \boldsymbol{X}_{it} + u_{2,it} \qquad (1)$$

Lagged renewal and maintenance costs are included in both equations to capture the dynamics in maintenance and renewals, as well as the interdependence between these activities.

Data

We have cost data for renewals and maintenance cost separately. That helps us build our two equation model in (1).

A full set of variables available for the analysis is given in Table E.1.

[1] Here we present the VAR(1) model for expositional simplicity. We consider further lags in the model estimation.

Table E.1 – Descriptive statistics, 1999-2014 (480 obs.)

	Mean	St.dev.	Min	Max
Hourly wage, SEK*	156.7	11.7	128.9	187.4
MaintC (Maintenance costs), million SEK*	56.78	44.37	8.03	334.41
RenwC (Renewal costs), million SEK*	40.74	63.95	0.00	452.13
Route length, km	280	174	13	989
Track length, km	358	229	39	1203
Length of switches, km	8.68	6.62	0.58	37.67
Length of structures (tunnels and bridges), km	5.72	7.22	0.55	40.43
Average age of rails	18.83	5.83	3.76	38.98
Ton-density (ton-km/route-km), million	7.9	7.2	0.2	33.2
Mixtend	0.06	0.24	0	1
Ctend	0.47	0.50	0	1
Trend	8.45	4.50	1	16

* 2014 prices.

Results

We find that past values of maintenance gives a better prediction of current renewal costs compared to only using past values of renewals as a predictor. That is, we find evidence for dynamic effects which is in turn an endorsement for our approach. Moreover, the results indicate intertemporal effects for both renewals and maintenance, where an increase in costs during a year predicts an increase in costs in the following year.

A particular purpose of estimating the dynamics in infrastructure costs is that it allows us to take these effects into account when assessing the cost impact of traffic. The estimated cost elasticity with respect to traffic is different in our dynamic model compared to static models that are frequently used in the literature on rail infrastructure costs. In particular, we used information on the dynamics between renewals and maintenance, in order to estimate equilibrium cost elasticities. These elasticities can be used in the calculation of marginal cost (which is the product of average costs and the cost elasticity), giving a better representation of the cost impact of an additional ton-km, considering that a traffic increase gives rise to costs in both the current year and subsequent years. Hence, the results in this paper are informative for infrastructure managers in Europe who need to set track access charges for the wear and tear caused by traffic.

Key elasticity results are shown in Table E.2. The parameter estimate for ton density in the maintenance equation is 0.2330 (p-value=0.010), which is in line with previous results on Swedish data (see for example Odolinski and Nilsson 2017 or Andersson 2008). In the renewal equation, the coefficient for ton density is 0.2633, which is lower than previous estimates on Swedish data (however, our estimate is not significantly different from zero, p-value = 0.498); Andersson et al. (2012) find a cost elasticity with respect to ton density at 0.547, and Yarmukhamedov et al. (2016) find elasticities between 0.5258 and 0.5646.

We calculate the equilibrium cost elasticities with respect to ton-density for both renewals and maintenance, using the results from model A2. These are presented in Table E.2, where γ^e denotes equilibrium cost elasticity. The elasticity for renewals is not significant at the 10 per cent level, while the estimates for maintenance are significant at the 1 per cent and 5 per cent level. All in all, the elasticities are larger than their static counterparts.

Table E.2 Key elasticity results from this study

	Cost elasticity	Coef.	Std. Err.
Dynamic Short run elasticity	$\gamma_{Maintenance}$	0.2330***	0.0901
	$\gamma_{Renewals}$	0.2633	0.3885
Dynamic Equilibrium elasticity	$\gamma^e_{Maintenance}$	0.3376**	0.1352
	$\gamma^e_{Renewals}$	0.3439	0.5173
Static comparator model	$\gamma_{Maintenance}$	0.2431***	0.0769
	$\gamma_{Renewals}$	-0.1189	0.4612
	$\gamma_{Main+Ren}$	0.2506**	0.1197

Notes: Short run elasticity is the change in cost resulting from a change in traffic in the same period; Equilibrium elasticity is the change in cost resulting from a change in traffic once all of the impact has traced through over time. Static comparator models represent results from more conventional contemporaneous models (they do not have dynamic terms included).

Overall, this work highlights that the dynamics in rail infrastructure costs are important to consider when setting track access charges with respect to the wear and tear caused by traffic.

The results can also be a useful demonstration of the maintenance and renewal strategy currently used. For example, the estimate for the second order lag of maintenance cost in the renewal equation gives us a hint on how sensitive renewal costs are to prior increases in maintenance. Moreover, the intertemporal effect for maintenance reveals how quickly this cost adjusts to equilibrium. Still, there is more to be done in this research area. For example, the analysis in this paper is not able to answer whether the quick adjustment in maintenance costs is avoiding an over-investment - that is, doing more than is necessary to uphold the performance of the infrastructure. In fact, the IM may well be over- or under-investing in maintenance after a sudden increase in traffic. User costs (values of train delays for passengers and freight companies) must be considered in this type of analysis. That is, with access to data on train delaying failures and delay costs for passengers and freight companies, it could be a step towards a cost-benefit analysis of maintenance and renewals which in turn can generate economically efficient levels of these activities. This is an area for future research.

Table of contents

Abbreviations and acronyms

Abbreviation / Acronym	Description
MC	Marginal cost
VAR	Vector Autoregressive Model

1. Introduction

The purpose of this paper is to provide empirical evidence on the wear and tear costs of rail infrastructure in Sweden, using a 16 year panel dataset. We consider both maintenance and renewals costs. To do this we consider a dynamic panel data specification which allows for interdependence between maintenance and renewals, as well as their intertemporal effects. The estimates can be used to calculate the marginal cost for traffic, which has become an important part of the track access charges that were introduced after the vertical separation between train operations and infrastructure management in Europe as of the 1990s. Given that there are dynamic effects between different activities in infrastructure provision, the marginal cost estimates that takes this into account will be closer to the actual cost of running one extra unit of traffic on the railway, compared to the cost estimates based on static models for maintenance (see for example Wheat et al. 2009) and renewals (see for example Andersson et al. 2012 and Andersson and Björklund 2012).

Knowledge on the intertemporal effects in rail maintenance and renewals and the interdependence between these costs is scarce. A notable exception is the study by Wheat (2015), in which a vector autoregressive model (VAR) is estimated for both maintenance and renewal costs in ten zones in Britain over a 15 year period. The study finds evidence on intertemporal effects, yet, not for a relationship between renewals and maintenance costs. An intertemporal effect is also found by Odolinski and Nilsson (2017) who estimate a dynamic model (system GMM) for maintenance. Similar to Wheat (2015), they find that an increase in maintenance costs in one year - due to for example a traffic increase – predicts an increase in maintenance costs in the next year. Other examples on research where the dynamics between maintenance and renewals are taken into account is Andersson (2008) and Odolinski and Smith (2016) who both use a dummy variable approach. However, it involves an arbitrary definition of major renewals and only allows for a stepwise effect of renewals on maintenance costs.

In this study we estimate a panel VAR model, which is an autoregressive model that considers several endogenous variables - renewals and maintenance in our case - in a multiple equation system. Our estimation approach is similar to Wheat's (2015), with the exception that we take the panel data structure into account. Hence, we are able to model unobserved individual heterogeneity, which in our case are contract specific effects. Moreover, we have access to ton-km instead of train-km, where the former provides a better representation of wear and tear.

The paper is organized as follows. In section 2, we describe the methodology used. It also includes a subsection on the equilibrium cost elasticity with respect to traffic; an elasticity that can be used in a calculation of the marginal cost for the wear and tear of the infrastructure. Section 3 comprises a description of the data. We specify our model in section 4, where we also present results from a test of the validity of our instruments and a stability test. The estimation results are presented in section 5. Section 6 concludes.

2. Methodology

Sims (1980) proposed the VAR model as an alternative to the simultaneous equation macroeconomic models prevalent at the time, which he criticized for its problems with arbitrary identification. The exogenous variables in the models - used for example to identify an effect on either the demand or supply - were often not strictly exogenous due to expectations in the economy that can change the behavior of the consumer (the demand) in addition to the variable's direct effect on the supplier and vice versa. Hence, there is a problem of simultaneity in the outcomes, which is the same type of problem we have with maintenance and renewals.

The objective with a VAR model is to capture the effects of unexpected exogenous shocks via identification strategies which, if properly specified, can make the model useful for forecasting and policy analysis. Thus, given the endogeneity of the maintenance and renewals described in the previous section, estimating a VAR model can be a fruitful approach for analyzing the dynamics in infrastructure provision.

We consider a panel VAR(p) model, where p denotes the lag length used in the model.[2] We have two endogenous variables: renewal costs (R_{it}) and maintenance costs (M_{it}), where $i = 1,2 \dots, N$ contract areas and $t = 1,2, \dots, T$ years. $\alpha_{1,i}$ and $\alpha_{2,i}$ are the unobserved individual-specific effects for the renewal and maintenance equations respectively, while $u_{1,it}$ and $u_{2,it}$ are their respective residuals, where $\left(u'_{1,it}, u_{2,it}\right) = \boldsymbol{u}_{it} \sim iid(0, \sum)$. $\sum$ is the covariance matrix of the errors. We also include a vector of exogenous variables $\boldsymbol{X}_{it}$ with parameters $\boldsymbol{\beta}_{11}$ and $\boldsymbol{\beta}_{21}$ for the maintenance and renewal equations respectively.

$$R_{it} = \alpha_{1,i} + \delta_{11} R_{it-1} + \theta_{11} M_{it-1} + \boldsymbol{\beta}_{11} \boldsymbol{X}_{it} + u_{1,it}$$

$$M_{it} = \alpha_{2,i} + \delta_{21} R_{it-1} + \theta_{21} M_{it-1} + \boldsymbol{\beta}_{21} \boldsymbol{X}_{it} + u_{2,it} \tag{1}$$

Lagged renewal and maintenance costs are included in both equations to capture the dynamics in maintenance and renewals, as well as the interdependence between these activities. We also consider a lagged traffic variable in the estimations.

We first perform model identification by making graphs of the data to spot trends and we also choose the lag order of the model based on model selection criteria. The lag order relates to autocorrelation in the residuals that can be removed by increasing the number of lags. For consistent estimation of the model parameters, $E\left(u'_{it}, u_{js}\right) = 0$, with $t \neq s$ i.e. no autocorrelation. The model check includes a stability test, which can reveal if a stationary process is generated by the model. If the process is non-stationary, first differencing would be required to avoid spurious results. However, the long-run relationship between the variables is eliminated when taking first differences. Moreover, if the variables are cointegrated (share a common trend), a vector-error correction model (VECM) is appropriate in order to analyze the long-run relationship between the variables (Heij et al. 2004). However, separating short-run and long-run relationships between maintenance and renewals is beyond the scope of this paper. Thus, we only consider the VAR model.

The model is estimated with generalized method of moments (GMM). In doing this, we need to consider that the lagged variables are correlated with the contract area specific effects. One way to deal with this problem is (again) to use first differencing, which removes these effects. However, first differencing is problematic for unbalanced panels (gap becomes larger). We therefore use the transformation proposed by Arrelano and Bover (1995), which is forward orthogonal deviation (or

[2] Here we present the VAR(1) model for expositional simplicity. We consider further lags in the model estimation.

Helmert transformation). More specifically, for each year and contract area, we subtract the mean of future observations.

We need to use instruments for the lagged variables, as these are correlated with the error terms. When including a set of lags as instruments, we use the method by Holz-Eakin et al. (1988), which basically substitutes missing values (created by increasing the lag length of the instruments) with zeros. This allows us to increase the lags of the instruments without losing the number of observations in the estimation.

2.1 Granger Causality

As a first test of interdependence between renewals and maintenance, we test whether lagged values of maintenance can improve the prediction of current values of renewals compared to only using lagged values of renewals (and vice versa). This approach of testing causal relations in time series is called a *Granger causality* test, proposed by Granger (1969).

A Granger causality test does however not reveal how exogenous changes in one variable affect another variable over time. To trace the effect of changes in renewals and maintenance costs, we make use of IRA, which requires identification of exogenous shocks ($\boldsymbol{\varepsilon}_{it}$). We assume these shocks to be a linear function of the residuals $\boldsymbol{u}_{it} = G\boldsymbol{\varepsilon}_{it}$, where G is a 2x2 matrix. Given our knowledge about the nature of renewals and maintenance, we choose *recursive identification* as the method to identify the shocks, which requires an ordering of the variables such that the G matrix can be calculated from the covariance matrix $\sum$ by using the Cholesky decomposition (see for example Sims 1989 and Christiano et al. 1999. For an overview of this method, see KVA 2011, p. 15-17). Simply put, the ordering should be constructed on the basis of how fast the variables respond, from slow to fast. In our case, renewals are ordered first as we assume that the only shock that can have an effect on current renewals is a shock in renewals, while current maintenance can be influenced by both a renewal shock and a maintenance shock. Notice that we only assume that a maintenance shock will not affect *current* renewals. Future renewals are likely to be influenced by maintenance shocks.

2.2 Equilibrium cost elasticity with respect to traffic

With lagged cost variables in our model, we are able to calculate the "equilibrium cost elasticity" with respect to traffic, both for renewals and maintenance.[3] For maintenance costs, the logic is that an increase in traffic in year $t-1$ will affect maintenance costs in year $t-1$, which in turn affects maintenance costs in year t. Moreover, maintenance costs that have adjusted into equilibrium imply $lnM_{it} = lnM_{it-1} = lnM_i^e$**.** The equation for maintenance costs in (1), with a logarithmic transformation and a traffic variable Q, is then

$$lnM_i^e = \alpha_{2,i} + \delta_{21} lnR_i^e + \theta_{21} lnM_i^e + \beta_{21} lnQ_i + \boldsymbol{\beta}_{23}\boldsymbol{X}_{it} + u_{2,i} \qquad (2)$$

[3] The term long-run cost is often used in the literature, which requires that there are no fixed factors in the production. However, in our analysis, the rail infrastructure is mainly fixed. We therefore prefer the term equilibrium costs.

which gives

$$lnM_i^e = \frac{\alpha_{2,i}}{1-\theta_{21}} + \frac{\delta_{21}}{1-\theta_{21}} lnR_i^e + \frac{\beta_{21}}{1-\theta_{21}} lnQ_i + \frac{\boldsymbol{\beta}_{22}}{1-\theta_{21}} \quad \boldsymbol{X}_{it} + \frac{u_{2,i}}{1-\theta_{21}} \tag{3}$$

The equilibrium cost elasticity with respect to traffic is then

$$\gamma_M^e = \frac{\partial lnM_i^e}{\partial lnQ_i} = \frac{\beta_{21}}{1-\theta_{21}} \tag{4}$$

The same logic applies for the renewal equation.

3. Data

Data has been obtained from the Swedish Transport Administration (the Infrastructure Manager; hereafter referred to as the IM), and consists of renewal and maintenance costs, traffic, and characteristics of the railway network such as track length and rail age. We also consider an input price variable (wages) in this study, which has been obtained from the Swedish Mediation Office (via Statistics Sweden). A complete list together with descriptive statistics is provided in Table 1 below.

Maintenance is activities performed to implement railway services according to the timetable and maintain the railway assets. As of 2007, snow removal is defined as maintenance and is included in the maintenance contracts. We are however able to pinpoint the snow removal costs in the data, and we exclude these costs due to its (stochastic) weather dependence. Renewals consist of replacements or refurbishments of the railway assets.

Maintenance and renewals are procured separately by the IM. The IM used in-house production of renewals until exposure to competition was introduced in 2001, while competitive tendering of maintenance services was introduced gradually in 2002.4 The effect competitive tendering had on renewal costs in Sweden has not been studied. However, in terms of maintenance costs, Odolinski and Smith (2016) find an 11 per cent reduction due to competitive tendering over the period 1999-2011. This indicates a structural change in infrastructure provision that needs to be considered when analyzing the interdependence between maintenance and renewals. Hence, Table 1 includes dummy variables for competitive tendering of railway maintenance; Mixtend which indicates the first year a contract area is tendered, as the first year is often a mix between not tendered and tendered in competition, and Ctend which indicates the subsequent years an area is tendered in competition. Tendering variables for renewals are not included in this study due to missing information.5

[4] Already in 1999, about 45 per cent of the reinvestment projects were produced by private companies (Trafikverket 2012).

[5] We do not think this is a significant issue for the estimation results as only one year in the estimation sample (2000) would include areas not tendered when using one lag in the model.

Data on infrastructure characteristics is available at a detailed level, while costs and traffic are reported at the more aggregate track section level. Moreover, each contract area for maintenance consists of several track sections. Considering that renewals can overlap adjacent track sections, we use contract areas as the identifier in our estimations. In that way, we have less artificial splits of renewal costs.

Table 1 – Descriptive statistics, 1999-2014 (480 obs.)

	Mean	St.dev.	Min	Max
Hourly wage, SEK*	156.7	11.7	128.9	187.4
MaintC (Maintenance costs), million SEK*	56.78	44.37	8.03	334.41
RenwC (Renewal costs), million SEK*	40.74	63.95	0.00	452.13
Route length, km	280	174	13	989
Track length, km	358	229	39	1203
Length of switches, km	8.68	6.62	0.58	37.67
Length of structures (tunnels and bridges), km	5.72	7.22	0.55	40.43
Average age of rails	18.83	5.83	3.76	38.98
Ton-density (ton-km/route-km), million	7.9	7.2	0.2	33.2
Mixtend	0.06	0.24	0	1
Ctend	0.47	0.50	0	1
Trend	8.45	4.50	1	16

* 2014 prices.

4. Model specification

To get a first impression of the main variables of interest, we make a graph of maintenance and renewal costs. We use costs per ton-km as we have an unbalanced panel of data. Both maintenance and renewal costs have an upward trend, yet renewal costs have a more lumpy nature with more variation during the studied period. Because in our dataset we aggregate to contract areas, the lumpy nature of renewals implies that our data have fewer observations with zero renewal costs compared to track sections (for example analyzed by Andersson et al (2012)).

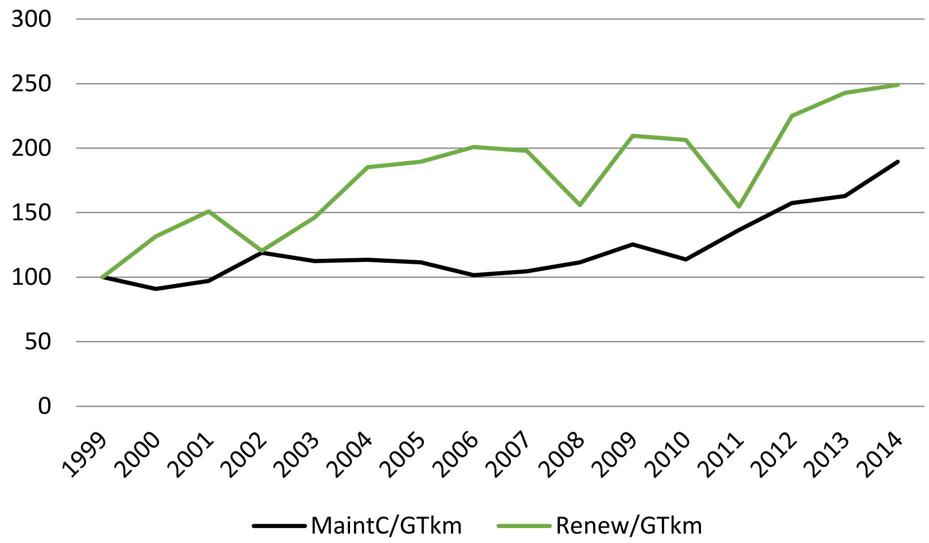


Figure 1 - Indices for maintenance (MaintC) and renewal costs (RenwC) per gross ton-km (GTkm), 1999-2014 (1999=100)

To control for fixed (time-invariant) effects in the variables, we time-demean the log transformed variables – that is, we subtract their group means: $\ddot{y}_{it} = y_{it} - \bar{y}_i$, where $\bar{y}_i = T^{-1}\sum_{t=1}^{T} y_{it}$. As previously noted, we also use a Helmert transformation in order to control for contract specific effects.

To determine the lag length of our model, we use the testing procedure proposed by Andrews and Lu (2001), which are consistent moment and model selection criteria (MMSC) versions of the Akaike information criterion (AIC), the Bayesian information criterion (BIC), and the Hannan-Quin information criterion (HQIC).[6] We start with the maximum number of lags and test down, where we choose the lag order with lowest values of the MMSC versions of AIC, BIC and HQIC. The test results show that the model with lag order 1 has the lowest values of AIC (-56.74), BIC (-262.57) and HQIC(-231.17) compared to the model with lag order 2, with the corresponding values -50.55, -234.62 and -208.55. However, we consider the second lag to be informative. We therefore also present results from models with lag order 2. Moreover, there is reason to consider a lagged traffic variable as an increase in traffic might trigger some maintenance activities in the next year. Note that this effect can be distinguished from the effect picked up by lagged maintenance costs, which determines the adjustment of costs caused by an increase in traffic and/or other cost drivers.

To test if our instruments are valid (if they are exogenous) we perform the Sargan test of overidentifying restrictions. The null hypothesis is that the instruments are valid. In our models in section 5, we fail to reject the null hypothesis and can conclude that our instruments are exogenous.

Finally, we test if the model estimations presented in the next section are stable. The modulus of each of the eigenvalues is below one, which implies that our estimated vector autoregressions are stable (Lütkepohl 2005, p.14-15).

[6] The formulations of the criteria in Abrigo and Love (2015) are used.

5. Results

Estimation results from Model A1 and A2 are presented in Table 2, where the former only includes the endogenous variables, and the latter includes a set of exogenous variables. The static comparison models (Models B1-B3) are presented in Table 3. The models are estimated with robust standard errors, using the iterative GMM estimator. We use the maximum lag length of the instruments (up to 14), which improves the efficiency of the model estimation; reducing the lag length to 12 or 11 for the instruments generates larger standard errors. All estimations are carried out with Stata 12 (StataCorp.2011) using the package provided by Abrigo and Love (2015).

The results for lagged maintenance and renewals are similar in both model A1 and A2 with respect to the signs of the coefficients for the lagged variables, yet the estimates for lagged maintenance in the maintenance equation are significantly lower when exogenous variables are included. This indicates that we may have omitted variable bias in Model A1. We focus on the results from Model A2 which includes variables for railway characteristics, ton density, dummy variables for competitive tendering and time trends.[7]

The significance tests of the parameter estimates for lagged variables in the maintenance and renewal equations can be interpreted as Granger causality tests. The prediction of current renewals is improved by lagged values of renewals, with a coefficient at 0.3193 that is significant at the 1 per cent level. This may seem odd, but a possible explanation is that budget restrictions can make it difficult to complete renewals of the railway assets during one year, which leaves some of the required renewals for the next year. Moreover, the coefficient for renewals costs in year t-2 predicts a decrease in current renewals. More specifically, the coefficient is -0.0843, yet with p-value=0.124. The estimated intertemporal effects for renewals then suggests that renewals within a contract area are likely to overlap between two years, and seem to have the expected decreasing effect on renewal costs in the subsequent year.

A lagged value of maintenance improves the prediction of current values of renewals compared to only using lagged values of renewals. The estimation results show that maintenance cost in year t-2 predicts an increase in renewals ($MaintC_{t-2}$ is 0.5776, with p-value 0.018). Hence, this model suggests that a shock in maintenance may increase a need for renewals in the second year, while it is unlikely to occur in the first year (coefficient is -0.1671 with p-value 0.540). The impact on renewals is rather intuitive considering that renewals should be preceded by large (corrective) maintenance costs as this is what generally motivates a renewal.

When it comes to lagged values of renewals in the maintenance equation, we do not find a significant Granger causality, and the estimate is close to zero. However, lagged maintenance costs predict an increase in current maintenance, with a coefficient at 0.3032 (p-value= 0.000). This estimate is somewhat higher than the coefficient in Odolinski and Nilsson (2017), who estimated a system GMM on Swedish data at the track section level (more observations available compared to the contract area level), generating a coefficient for lagged maintenance costs at 0.2140.

Traffic is a key driver of cost. Therefore, the cost elasticities with respect to ton density are of particular interest which, together with coefficients for lagged costs, allows us to estimate equilibrium cost elasticities. The parameter estimate for ton density in the maintenance equation is 0.2330 (p-value=0.010), which is in line with previous results on Swedish data (see for example Odolinski and

[7] We tried different restrictions with respect to the variables for railway characteristics. This did not have a significant effect on the results.

Nilsson 2017 or Andersson 2008). In the renewal equation, the coefficient for ton density is 0.2633, which is lower than previous estimates on Swedish data (however, our estimate is not significantly different from zero, p-value = 0.498); Andersson et al. (2012) find a cost elasticity with respect to ton density at 0.547, and Yarmukhamedov et al. (2016) find elasticities between 0.5258 and 0.5646.

For comparison, we estimate the static counterparts of the models, including a model with the sum of maintenance and renewal costs as the dependent variable (see Table 3). The renewal model (B1) generates non-satisfactory results due to a negative and insignificant traffic elasticity estimate, which is not surprising given the lumpy nature of renewals. The results in models B2 and B3 are more in line with the maintenance equation results in model A2, with similar cost elasticities with respect to ton density even though renewals are included in model B3.

Table 2 Estimation results, Models A1 and A2, with order 2 lags (342 obs.)

		Model A1		*Model A2*	
Equation	Variable	Coef.	Std. Err.	Coef.	Std. Err.
RenwC	RenwC_t-1	0.3361***	0.0655	0.3193***	0.0678
	RenwC_t-2	-0.0718	0.0521	-0.0848	0.0544
	MaintC_t-1	-0.3918	0.2645	-0.1671	0.2729
	MaintC_t-2	0.2535	0.2829	0.5776**	0.2436
	Ton density	-	-	0.2633	0.3885
	Track length	-	-	2.2840	1.6545
	Rail age	-	-	-0.0591	0.6446
	Switch length	-	-	-0.6738	0.9885
	Length of struct.	-	-	0.6425	0.6824
	Trend	-	-	0.0263	0.2142
	Trend^2	-	-	-0.0104	0.0199
	Mixtend	-	-	0.3549	0.3445
	Ctend	-	-	0.0734	0.3759
MaintC	RenwC_t-1	-0.0107	0.0099	0.0044	0.0106
	RenwC_t-2	-0.0021	0.0094	0.0091	0.0093
	MaintC_t-1	0.4665***	0.0570	0.3032***	0.0560
	MaintC_t-2	0.1530***	0.0522	0.0065	0.0503
	Ton density	-	-	0.2330***	0.0901
	Track length	-	-	0.2181	0.2151
	Rail age	-	-	0.1617	0.1109
	Switch length	-	-	0.3765***	0.1331
	Length of struct.	-	-	0.3852***	0.1323
	Trend	-	-	-0.1339***	0.0325
	Trend^2	-	-	0.0157***	0.0030
	Mixtend	-	-	-0.0694	0.0639
	Ctend	-	-	-0.1145*	0.0695
Moduli of eigenvalues		0.69, 0.26, 0.26 and 0.22		0.42, 0.35, 0.35 and 0.11	

***, **, * : Significance at the 1%, 5%, and 10% level, respectively.

Table 3 – Estimation results, Models B1-B3 (342 obs.)

	Model B1		*Model B2*		*Model B3*	
Dependent variable	RenwC		MaintC		RenwC+MaintC	
	Coef.	Std. Err.	Coef.	Std. Err.	Coef.	Std. Err.
Ton density	-0.1189	0.4612	0.2431***	0.0769	0.2506**	0.1197
Track length	3.2393**	1.5187	0.5908***	0.2232	0.6333**	0.3236
Rail age	-0.2836	0.4919	0.1754	0.1139	0.2027	0.1658
Switch length	-1.4351**	0.6651	0.4559***	0.1532	0.2406	0.2149
Length of structures	0.2130	0.6583	0.3610***	0.1191	0.3627*	0.1953
Trend	0.2501	0.2382	-0.1588***	0.0367	-0.0994	0.0709
Trend^2	-0.0323	0.0219	0.0190***	0.0033	0.0117*	0.0064
Mixtend	0.2823	0.4257	0.1019	0.0775	0.1264	0.1456
Ctend	0.1988	0.4526	-0.0051	0.0818	0.0518	0.1502

***, **, *: Significance at the 1%, 5%, and 10% level, respectively.

We calculate the equilibrium cost elasticities with respect to ton-density for both renewals and maintenance, using the results from model A2. These are presented in Table 4, where γ^e denotes equilibrium cost elasticity. The elasticity for renewals is not significant at the 10 per cent level, while the estimates for maintenance are significant at the 1 per cent level. All in all, the elasticities are larger than their static counterparts.

Table 4 – Equilibrium cost elasticities with respect to ton density, Model A2

Cost elasticity	Coef.	Std. Err.
$\gamma^e_{Maintenance}$	0.3376**	0.1352
$\gamma^e_{Renewals}$	0.3439	0.5173

***, **, *: Significance at the 1%, 5%, and 10% level, respectively.

The dummy variable for tendering of maintenance contracts shows that maintenance costs decreased with about 11 per cent[8], similar to the results in Odolinski and Smith (2016). In the renewal equation, the estimate for competitive tendering of maintenance is not significantly different from zero. In one way, this is not surprising considering that a decision to renew is not likely to be directly connected to the introduction of tendering of maintenance; the decision to renew ought to be more connected to the condition of the railway assets and how costly infrastructure failures are for society on a certain part of the network. However, the amount and/or type of maintenance carried out - which may have

[8] exp(-0.1145)-1 = -0.1082.

changed due to competitive tendering - is certainly connected to the condition of the railway assets, which affects the need for renewals. Still, as previously noted, the results do not indicate that competitive tendering of maintenance has affected the renewal costs. It should also be noted that a difference-in-differences approach would be a more accurate way of estimating the effect of tendering, an approach used in Odolinski and Smith (2016).

Finally, we note that the estimates for track length, average rail age, and length of structures have the expected signs in the maintenance equation. The estimate for average rail age is close to zero in the renewal equation, and not statistically significant, as is the coefficient for switch length.[9] It is only the coefficients for switches and structures in the maintenance equation that is statistically significant. Using lag order 1 does not change these coefficients in the maintenance equation significantly.

6. Conclusions

In this paper we have estimated a panel VAR model on rail infrastructure costs in Sweden. The results provide empirical evidence on the relationship between maintenance and renewals, as well as evidence on intertemporal effects for each of these activities. The results show that past maintenance costs can improve the prediction of current values of renewals compared to only using past values of renewals. We also found intertemporal effects for both renewal and maintenance costs; an increase in renewals (maintenance) during one year predicts an increase in renewals (maintenance) during the next.

A particular purpose of estimating the dynamics in infrastructure costs is that it allows us to take these effects into account when assessing the cost impact of traffic. The estimated cost elasticity with respect to traffic is different in our dynamic model compared to static models that are frequently used in the literature on rail infrastructure costs. In particular, we used information on the dynamics between renewals and maintenance, in order to estimate equilibrium cost elasticities. These elasticities can be used in the calculation of marginal cost (which is the product of average costs and the cost elasticity), giving a better representation of the cost impact of an additional ton-km, considering that a traffic increase gives rise to costs in both the current year and subsequent years. Hence, the results in this paper are informative for infrastructure managers in Europe who need to set track access charges for the wear and tear caused by traffic.

The results can also be a useful demonstration of the maintenance and renewal strategy currently used. For example, the estimate for the second order lag of maintenance cost in the renewal equation gives us a hint on how sensitive renewal costs are to prior increases in maintenance. Moreover, the intertemporal effect for maintenance reveals how quickly this cost adjusts to equilibrium. Still, there is more to be done in this research area. For example, the analysis in this paper is not able to answer whether the quick adjustment in maintenance costs is avoiding an over-investment - that is, doing more than is necessary to uphold the performance of the infrastructure. In fact, the IM may well be over- or under-investing in maintenance after a sudden increase in traffic. User costs (values of train delays for passengers and freight companies) must be considered in this type of analysis. That is, with access to data on train delaying failures and delay costs for passengers and freight companies, it could be a step towards a cost-benefit analysis of maintenance and renewals which in turn can generate economically efficient levels of these activities. This is an area for future research.

[9] We also estimated the model without switch length, which did not affect the results significantly.

7. References

Abrigo, M.R.M., and I. Love (2015): 'Estimation of Panel Vector Autoregression in Stata: a Package of Programs', Working paper, May 2015.

Andersson, M. (2008): 'Marginal Railway Infrastructure Costs in a Dynamic Context', *EJTIR*, 8, 268-286.

Andersson, M. and G. Björklund (2012): 'Marginal Railway Track Renewal Costs: A Survival Data Approach', CTS Working Paper 2012:29, Centre for Transport Studies, Stockholm.

Andersson, M., A. Smith, Å. Wikberg, and P. Wheat (2012): 'Estimating the marginal cost of railway track renewals using corner solution models', *Transportation Research Part A*, 46, 954-964.

Andrews, D.W.K, and B. Lu (2001): 'Consistent model and moment selection procedures for GMM estimation with application to dynamic panel data models', *Journal of Econometrics*, 101, 123-164.

Arellano, M., and O. Bover (1995): 'Another look at the instrumental variable estimation of error-components models', *Journal of Econometrics*, 68, 29-51

Christiano, L.J., M. Eichenbaum, and C.L. Evans (1999): 'Monetary Policy Shocks: What Have We Learned and to What End?', In: Taylor, J.B. and M. Woodford (Eds), *Handbook of Macroeconomics*, Volume 1A, Elsevier Science, North-Holland, 65-148.

Granger, C.W.J. (1969): 'Investigating Causal Relations by Econometric Models and Cross-spectral Methods', *Econometrica*, 37(3), 424-438.

Heij, C., P. de Boer, P.H. Franses, T. Kloek, and H.K. van Dijk (2004): 'Econometric Methods with Applications in Business and Economics', Oxford University Press Inc., New York.

Holtz-Eakin, D., W. Newey, and H.S. Rosen (1988): 'Estimating vector autoregressions with panel data', *Econometrica,* 56(6), 1371-1395.

KVA (2011): 'Empirical Macroeconomics', Scientific Background on the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2011 compiled by the Economic Sciences Price Committee of the Royal Swedish Academy of Sciences. Kungliga Vetenskapsakademien, 10 October 2011.

Lütkepohl, H. (2005): 'New Introduction to Multiple Time Series Analysis', Springer-Verlag Berlin, Heidelberg, 2005.

Odolinski, K. and A.S.J. Smith (2016): 'Assessing the Cost Impact of Competitive Tendering in Rail Infrastructure Maintenance Services: Evidence from the Swedish Reforms (1999 to 2011)', *Journal of Transport Economics and Policy*, 50(1), 93-112.

Odolinski, K. and J-E. Nilsson (2017): 'Estimating the marginal maintenance cost of rail infrastructure usage in Sweden; does more data make a difference?', *Economics of Transportation*, 10, 8-17.

Sims, C.A. (1980): 'Macroeconomics and reality', *Econometrica*, 48, 1-48.

Sims, C.A. (1989): 'Models and Their Uses", *American Journal of Agricultural Economics*, 71, 489-494.

StataCorp.2011. *Stata Statistical Software: Release 12*. College Station, TX: StataCorp LP.

Trafikverket (2012): 'Organiseringen av underhåll av den svenska järnvägsinfrastrukturen', Trafikverket Rapport, Dnr 2012/63556. (In Swedish)

Wheat, P., A.S.J. Smith, and C. Nash (2009): 'CATRIN (Cost Allocation of TRansport INfrastructure cost)', Deliverable 8 – Rail Cost Allocation for Europe, VTI, Stockholm.

Wheat, P. (2015): 'The sustainable freight railway: Designing the freight vehicle–track system for higher delivered tonnage with improved availability at reduced cost SUSTRAIL', Deliverable 5.3: access charge final report Annex 4, British Case Study.

Yarmukhamedov, S., J-E. Nilsson, K. Odolinski (2016): 'The marginal cost of reinvestments in Sweden's railway network', VTI notat 23A-2016.