FACULTY OF ECONOMICS AND
BUSINESS ADMINISTRATION

RAILWAY TRAFFIC CONTROL PERFORMANCE

AN EMPIRICAL ANALYSIS OF PRODUCTIVE EFFICIENCY, FATIGUE RISK, AND HUMAN ERROR IN A REAL-WORLD SETTING

Bart Roets

Supervisor: Prof. Dr. Johan Christiaens

A dissertation submitted to Ghent University in partial fulfilment of the requirements for the degree of Doctor in Business Economics

Academic year: 2017 – 2018

GHENT
UNIVERSITY

DOCTORAL ADVISORY COMMITTEE

Prof. Dr. Johan Christiaens

Supervisor, Ghent University

Prof. Dr. Mario Vanhoucke

Ghent University & Vlerick Business School & University College London (UK)

Dr. Sabine Verboven

Hasselt University & Atlas Copco

Prof. Dr. Joris Voets

Ghent University

DOCTORAL EXAMINATION BOARD

Prof. Dr. Patrick Van Kenhove

Dean of the Faculty of Economics and Business Administration, Ghent University

Prof. Dr. Joris Voets

Academic Secretary, Ghent University

Prof. Dr. Johan Christiaens

Supervisor, Ghent University

Prof. Dr. Simon Folkard

Swansea University (UK) & Université Paris Descartes (France)

Prof. Dr. Konstantinos P. Triantis

Virginia Polytechnic Institute and State University (USA)

Prof. Dr. Mario Vanhoucke

Ghent University & Vlerick Business School & University College London (UK)

Dr. Sabine Verboven

Hasselt University & Atlas Copco

PUBLICATION STATUS

First paper: published, European Journal of Transport and Infrastructure Research (Web of Science), 2015. Title: 'Evaluation of railway traffic control efficiency and its determinants'.

Second paper: published online, Journal of Transportation Safety & Security (Web of Science). Title: 'Shift work, fatigue and human error: an empirical analysis of railway traffic control'.

Third paper: second review round, European Journal of Operational Research (Web of Science). Title: 'Multi-output efficiency and human error: an analysis of railway traffic control centre performance'.

Contents

List of Figures

List of Tables

SAMENVATTING

De Europese spoorwegen ervaren toenemende druk om hun efficiëntie te verhogen zonder de veiligheid in het gedrang te brengen. Ter ondersteuning van deze uitdaging onderzoekt dit proefschrift de verkeersleiding, een kernactiviteit van het spoor die sterk leunt op efficiëntie en veiligheid om haar performantie te verhogen.

De gebundelde artikels presenteren een aantal geavanceerde instrumenten, empirische resultaten, en beleidsaanbevelingen die geworteld zijn in twee uiteenlopende wetenschappelijke disciplines: efficiëntiemeting en modellering van vermoeidheidsrisico's. Het multidisciplinaire karakter van het proefschrift laat een breder, niet-unidimensioneel antwoord toe op de overkoepelende onderzoeksvraag: hoe de personeelsbezetting efficiënter krijgen, zonder hierbij menselijke factoren en veiligheid uit het oog te verliezen.

Door deze aanpak introduceert dit doctoraat het nieuwe onderzoeksdomein van 'efficiëntiemeting in spoorverkeersleiding', verschaft het diepgaander inzicht in de erg schaars onderzochte vermoeidheidseffecten bij spoorverkeersleiders, en lieert het de kans op menselijke fouten met zowel efficiëntie- als vermoeidheidsresultaten.

De ontwikkelde modellen en de empirische bevindingen worden ondersteund door een op maat gebouwde Business Intelligence toepassing, rechtstreeks gevoed door Belgische spoorwegdata. Dit overbrugt de kloof tussen onderzoekers en spoorexperten op actieve wijze, en verhoogt hierdoor de indruksvaliditeit van het onderzoek op substantiële wijze.

SUMMARY

European railways are under increasing pressure to raise efficiency without sacrificing safety. In support of these challenges, this dissertation aims to assess railway traffic control, a core railway activity which leans heavily on efficiency and safety to improve its performance.

The presented research offers a series of advanced tools, empirical findings, and policy recommendations rooted in two distinct disciplines: efficiency estimation and fatigue risk modelling. The multidisciplinary nature of the dissertation allows for a broader, non-unidimensional answer to the overarching question which has driven this research: how to improve staffing efficiency, while accounting for human factors and safety concerns.

As such, the dissertation introduces the new field of railway traffic control efficiency, deepens insight in the severely underresearched area of railway traffic controller fatigue, and links the probability of human error to both staffing efficiency and fatigue risk.

The developed models and the empirical findings are supported by a purpose-built Business Intelligence environment, fuelled by real-world Belgian railway data. This actively bridges the gap between researchers and railway experts, and as such substantially leverages the face-validity of the research.

Chapter 1

RESEARCH GAPS AND CONTRIBUTIONS

1.1 Research gaps

This dissertation aims to assess **railway traffic control performance**, from **different and possibly counterbalancing perspectives**. Using unique intra-company data and expert knowledge from Infrabel, the Belgian railway infrastructure operator, it develops a series of performance analysis models and tools. As such, it grasps the opportunities offered by the continuing digitization of Europe's traffic control systems, and presents data-driven empirical research that includes concerns about efficiency, fatigue, safety, and human error.

Two performance perspectives are analysed. A **first performance component** of interest is the **productive efficiency** of the railway traffic control process, a thus far unexplored area in the efficiency and transportation literature. Since 1991, European directives gradually unbundled the railway system into national 'infrastructure managers', and several competing railway undertakings. This vertical separation of infrastructure and train operations, one of the cornerstones of Europe's railway policy, has increased the academic attention towards the cost and efficiency of railway infrastructure. The existing body of literature in this research area has been steadily complemented by specific infrastructure oriented research, with a main focus on marginal cost estimation (e.g. Johansson and Nilsson, 2004; Wheat and Smith, 2008; Andersson, 2008) and efficiency measurement (e.g. Kennedy and Smith, 2004; Smith et al., 2010; Smith, 2012. As such, the focus of this **previous research was limited to asset management** (building and maintaining the network), and was almost exclusively based on parametric techniques. Although progressively emerging in railway industry reports, the

efficiency of railway traffic control (i.e. 'the second component of operating the infrastructure'; Cowie and Loynes, 2012) **consistently remained out of scope** of all previous efficiency research. As such, and although a strand of air traffic control efficiency studies is gradually taking shape (e.g. Button and Neiva, 2014; Bilotkach et al., 2015), there is no previous research on railway traffic control efficiency.

With this dissertation we therefore **fill a gap in the efficiency and transportation literature**, and initiate a new research field with promising potential for empirical analysis and real-life implementation. In order to assess efficiency, we customise the **Data Envelopment Analysis (DEA) methodology** and evaluate **staffing efficiency in railway traffic control centres**. DEA is a well-established tool for measuring the efficiency of Decision Making Units (DMUs), which convert multiple inputs into multiple outputs (Charnes et al., 1978). Relying on Linear Programming techniques, the DEA methodology has sparked a large number of empirical efficiency studies in the past decades (for an overview, see the collections of e.g. Fried et al., 2008; Cooper et al., 2011). The distinguishing feature of the methodology is its non-parametric nature, implying that no a priori (typically unverifiable) functional form specifications are imposed on the production technology. As such, DEA is a data-oriented managerial decision-support tool.

To capture the traffic control process in a realistic way, we **tailor the DEA methodology to fit the situation at hand**. First, for non-computerised traffic control centres, we apply the Banker and Morey (1986) DEA model with a categorical variable, and extend the subsample bootstrap algorithm (Kneip et al., 2008) to account for the (non-convexity) properties of the Banker and Morey model. Second, for computerised traffic control centres, we customize and extend the Cherchye et al. (2013) multi-output efficiency measurement model. In conventional DEA, the production unit under consideration (e.g. the traffic control centre) is modelled as a 'black box' which transforms the inputs into outputs. The Cherchye et al. (2013) model incorporates information on the internal production process, which 'opens the black box' and provides efficiency results which can deliver additional managerial insights (for an overview on modelling internal production structures in DEA see Castelli et al., 2010). We formally include a priori information on the relation between inputs and outputs in the Cherchye et al. (2013) model, and demonstrate the transparency and flexibility of the ap-

proach. Finally, by statistically relating the obtained multi-output efficiency results with observations of human error, our contribution to the literature **extends beyond the scope of traditional efficiency methods**.

The **second performance component** is, in contrast with the rather Tayloristic efficiency perspective described above, oriented towards human behaviour and its consequences. In general, the rise of the 24-hour society is leading to an ever-increasing demand for round-the-clock shift work. The impact of this shift work on safety and performance, as well as on employee health and well-being, has been extensively studied (see, e.g., the research synthesis by Tucker et al., 2012). Driven by (public) safety concerns, this dissertation focuses on **railway traffic controller fatigue** and its **impact on human error**. Despite their safety-critical role and the recognized fatigue levels in practice (Gertler et al., 2013; Rail Safety and Standards Board, 2015), railway traffic controller fatigue remains a **severely underresearched area**. Fatigue studies in the railway industry have mainly focused on train drivers, and there are only a limited (but growing) number of sleep and fatigue publications involving railway traffic controllers (e.g., Popkin et al., 2001; Dorrian et al., 2011; Cotrim et al., 2017).

Fatigue is recognised as a **major contributing factor to transportation accidents**. It is described by Åkerstedt (2000) as 'the largest identifiable and preventable cause of accidents in transport operations', causing an estimated 15 to 20% of all accidents. Even before drowsiness or involuntary micro-sleeps set in, fatigue impairs cognitive performance and as such causes memory lapses, decreases attention, or lowers reaction time. Providing an almost palpable sense of fatigue-induced performance impairment, a seminal experiment by Dawson and Reid (1997) reveals that even moderate levels of extended wakefulness can lead to a performance degradation equivalent to alcohol intoxication: after 17 hours of extended wakefulness, the observed cognitive psychomotor performance (hand-eye coordination) reaches a level corresponding to a 0.05% blood alcohol concentration, while after 24 hours the performance decrement equals a 0.10% intoxication.

Shiftworkers, like the railway traffic controller population under study, are particularly prone to fatigue and its consequences. In order to mitigate fatigue-related safety issues, **organizations increasingly rely on fatigue risk or 'biomathematical' models** to predict the fatigue or risk associated with shift work (Gander et al., 2011; Darwent et al.,

2015). Fatigue risk models can be categorized in two main groups (Dawson et al., 2011). The so-called one step models directly apply sleep-wake data, such as work and sleep diaries and/or wrist actigraph registrations, as an immediate input for fatigue estimations. Two-step models first evaluate a given work schedule to estimate an average sleep-wake pattern, and subsequently predict fatigue levels on the basis of these estimations. Two-step models have a distinct practical advantage in the real world, as staff schedules are readily available information in workplace settings (Fletcher and Dawson, 2001; Dean et al., 2007). However, the two-step estimation procedure induces additional variance in the fatigue risk estimations, which can lead to a decrease in its predictive ability. Nonetheless, there has been little research on the statistical reliability of two-step models, which can **cast doubts on their validity in real-world settings** (Dawson et al., 2011). In addition, current fatigue risk models are mainly based on biological determinants of fatigue, and **fail to sufficiently address the influence of psychosocial factors on sleep-wake behaviour** (ibid.). For example, sleep duration and quality can be negatively impacted when the need for recovery sleep competes with social demands or family constraints. These findings correlate with a recent report from the UK railway industry (Rail Safety and Standards Board, 2015), in which not only work-related factors but also home-life related activities were cited as the most common fatigue sources.

In response to these fatigue risk modelling deficiencies reported in the literature, and for the purpose of our railway traffic control research, we evaluate the **predictive validity** of a commonly used work schedule-based fatigue risk tool: the Folkard et al. (2007) **Risk Index**. Only one preceding study has examined the validity of the Risk Index (Greubel and Nachreiner, 2013), and as such we extend this previous (internet survey-based) research to real-world settings. In addition, similar to the traffic control efficiency analysis, we relate the fatigue risk predictions with the probability of human error. We also investigate the effect of additional risk factors on human error probability (such as the day of the week, which can induce social and family-related pressure on sleep and recovery). As such, this dissertation not only **contributes** to the **underresearched area of railway traffic controller fatigue**, but **also to the fatigue modelling literature**.

1.2 Specific contributions of each chapter

1.2.1 Structure of the dissertation

Table 1.1 highlights the contributions of the three articles presented in this dissertation. The efficiency benchmarking model presented in chapter 2 is developed for **non-computerized traffic control centres**, while the 3rd and 4th chapter are oriented towards **digitized traffic control**. The rich micro-data provided by these traffic control centres allows for an in-depth analysis of fatigue risk (chapter 3) and staffing efficiency (chapter 4), while explicitly considering the relationship with human error. **Each chapter progressively increases the level of data disaggregation**: chapter 2 is based on 900 monthly observations, chapter 3 on 11,000 work shifts, and chapter 4 on 83,000 hourly observations.

As indicated above, the contributions of each chapter stretch beyond the mere scope of railway traffic control. For example, chapter 3 also contributes to the fatigue modelling literature, by evidencing a significant association between the day of the week and the probability of human error. As a another example, chapter 4 presents the first efficiency analysis with an hourly time resolution, and demonstrates how this exceptionally disaggregated analysis allows to reveal staff schedule inefficiencies. The next sections discuss the contribution of each chapter more in detail. General conclusions and suggestions for further research are set out in the final chapter.

Table 1.1: Contributions of each chapter

Chapter 2: Railway traffic control efficiency and its determinants
Presents first efficiency analysis of railway traffic control
Presents a categorical variable Data Envelopment Analysis model with subsample bootstrap
Second-stage regressions reveal impact of policy-relevant factors (e.g. lean infrastructure)
Chapter 3: Shift work, fatigue risk and human error
Presents first real-word validation of the Risk Index (work schedule-based prediction of fatigue risk)
Contributes to the underresearched area of railway traffic controller fatigue
Tobit regressions reveal relation between probability of human error and day-of-the-week
Chapter 4: Multi-output efficiency and human error
Presents first efficiency analysis of computerised railway traffic control
Presents first hourly efficiency measurement of 24/7 services
Presents a multi-output Data Envelopment Analysis-based model with formal cost allocations
Probit regressions reveal relation between probability of human error and tasks with variable workload

1.2.2 Chapter 2: Railway traffic control efficiency and its determinants

This chapter **identifies and fills a gap in the literature**, by being the first to examine the productive efficiency of railway traffic control. Across Europe, railway traffic control technology is currently migrating from non-computerized (legacy-based) systems towards a digitized and centralized environment (Wilson and Norris, 2005). A benchmarking report published by the UK Office of Rail Regulation provides an international overview of this long-term technological shift (Civity management consultants, 2013). With 7 European national railway infrastructure managers participating in the study, the report states that the levels of traffic control centralisation and automation vary significantly across Europe. It identifies a series of leaders with a high degree of centralisation (such as the Dutch infrastructure manager ProRail) and followers, fully progressing in ambitious modernisation projects (e.g. Network Rail in the UK, Réseau Ferré de France, or Infrabel in Belgium). It is clear however that, despite these large-scale and long term investment projects, railway traffic control currently still remains a labour-intensive process in many European countries.

We therefore **develop a two-stage benchmarking framework**, which assesses and explains the staffing efficiency of **non-computerized traffic control centres**. In the first stage, a DEA model assesses monthly efficiency, and closely monitors performance trends over time. To circumvent issues of data availability, related to the non-digitized environment, we model one of the outputs as an (ordered) categorical variable. In the second stage, a series of regression models examine the impact of several exogenous and policy-related factors on productive efficiency. Before engaging in the second stage regression analysis, we lean on the subsample bootstrap algorithm (Kneip et al., 2008) to obtain bias-corrected efficiency estimates. We adapt the bootstrapping algorithm to accommodate for the categorical variable in the DEA model. As such, this paper also presents a subsample bootstrap algorithm applicable to the Banker and Morey (1986) DEA models. The second stage performs an OLS regression on the bias-corrected efficiency estimates, with separate models for weekday and weekend data. As a methodological robustness check, we complement the approach with a truncated regression, applying the single bootstrap procedure developed by Simar and Wilson (2007).

The proposed framework **can be adopted by railway infrastructure managers as an internal benchmarking tool**, evaluating the entire network or specific sub-regions. The single overall measure of efficiency obtained through the DEA calculations can act as a guide to pinpoint the best, good and worst practices throughout the examined area. Especially for large networks with an extensive number of non-computerized traffic control centres (such as the French or British, or German) this can deliver powerful management insight. The practical applicability of the approach is demonstrated with a unique 18-month dataset of Belgian relay technology traffic control centres. Aiming to uncover additional patterns and insights, we perform our analysis on two subsets of the monthly data: one covering the working week, the other the weekends.

The second-stage results suggest that in order to improve on traffic control efficiency, infrastructure managers should aim for **geographical concentration, larger team sizes, and a continuous follow-up of control centre opening times**. Further efficiency gains can be generated by reducing infrastructure complexity. This implies that an **asset management strategy**, aiming for 'lean infrastructure' (International Union of Railways (2002)

InfraCost study), is not only reducing maintenance cost, but also has **positive effects on traffic control efficiency**. In line with the overarching research question of this dissertation, we also examine the relationship between efficiency and human errors (number of train delays caused by traffic controller mistakes), but no significant association is found. Finally, the results also indicate that management should take into account the **differences between working week and weekend** when measuring and analysing traffic control performance. Although mean weekend efficiency is significantly lower than working week efficiency, a substantial number of traffic control centres exhibit a higher efficiency during the weekends. These diverging 'weekend effects' can further assist decision makers in identifying and analysing best and good practices, which may be different in weekends compared to the working week.

1.2.3 Chapter 3: Shift work, fatigue risk and human error

With this chapter, we focus on the human factor component of railway traffic control performance, and more specifically **fatigue risk and its link with human error**. We evaluate the predictive validity of a work schedule-based fatigue risk tool (the Folkard et al. (2007) **Risk Index**), and investigate the effect of additional risk factors (age, gender, part-time work, and day-of-week) on human error probability.

The original Risk Index was developed in Folkard and Lombardi (2004, 2006) and was further extended in Folkard et al. (2007). The Risk Index holds the specific advantage that it is primarily based on the risk of an accident or incident, whereas conventional fatigue models hinge on fatigue or alertness levels. The link between fatigue and safety is complex and does not always follow a linear path (Williamson et al., 2011), and can be influenced by a number of factors. As an example, Cabon et al. (2012) point at 'fatigue awareness', which can trigger fatigued staff to adopt risk mitigating strategies, such as an intentional increase of automation levels (e.g. a higher reliance on autopilot by tired flight crews).

Although the Risk Index is strictly speaking not a 'two-step' fatigue risk model (see the previous section), we categorize it as such for the purpose of this validation study, as (i) it directly relies on the input of work schedules and (ii) partly and indirectly leans on sleep restriction research (e.g. the Belenky et al. (2003) laboratory study on recovery sleep). As indicated by Dawson et al. (2011), there is a **lack of testing and validation studies for**

work schedule-based (two-step) models of fatigue.

By **linking workforce and operational data**, we are able to generate datasets which contain the necessary information at a highly disaggregate level (more than 11,000 work shifts are examined, including data on human error occurrence). The human errors considered consist of relatively frequent but non safety-critical task errors, detected by the computerized traffic control system and subsequently archived for analytical purposes. Categorizable as attention failures, the human errors can indirectly impact safety in railway transportation. At present, the human errors cannot be directly linked to the individual traffic controller. Therefore, we calculate the total error frequency for the entire traffic control team present during the work shift. We control for exposure by dividing the error frequency by the traffic volume, and as such obtain a transparent error rate of 'number errors per 1000 train movements'.

We apply bivariate correlations to evaluate the predictive validity of the Risk Index, and multivariate regressions to analyse the impact of the additional risk contributing factors. The regression model also takes into account the specific operational conditions of the railway traffic control environment, by controlling for local fixed effects and the level of traffic control automation. We account for the 11% of observations (i.e., work shifts) with zero error occurrence through a Tobit regression model for corner solution data. In order to test the predictive accuracy of the regression models, we perform a 5-fold cross-validation procedure.

The correlation results **validate the Risk Index in a railway traffic control setting**. This extends previous research on the validity of work schedule-based (two-step) fatigue risk models in general, and the Risk Index in particular. With specific regard to Risk Index, our results augment the previous internet survey-based research by Greubel and Nachreiner (2013) to real-world settings[1]. As such, we provide the first validation study of the Risk Index in an operational environment, more particularly in the railway industry, where the model is widely used (UK Rail Safety and Standards Board, 2015).

The Tobit regression results not only corroborate the correlation-based validation, but also **reveal risk predictors above and beyond shift schedule design**: consistently significant

[1]Greubel and Nachreiner (2013) also call for a more systematic analysis, based on larger samples, in order to further develop the Risk Index.

day-of-week effects are observed, during and surrounding the weekend. By decomposing the Tobit marginal effects in probability and size effects, the paper further deepens empirical and managerial insight. The probability of having at least one error is highest on Saturdays (a 6 percentage point higher probability compared to Mondays), and lowest on Tuesday, Wednesday and Thursday (around 2 percentage point decrease in probability). Beyond its empirical value, this quantitative information can be easily conveyed to non-statisticians, and as such can support management policy and communication with actionable language. In general, the results suggest that **safe work schedule design should also take into account the day of the week**, and not exclusively rely on current fatigue model outputs.

The significant 'day-of-week effects' are a clear demonstration of the **current lack of fatigue risk models to account for psychosocial determinants** (such as social demands during non-work periods, Di Milia et al., 2011). This not only relates to the investigated Risk Index, but also to the scant body of fatigue research on day-of-week effects (e.g., Monk and Wagner, 1989; Brogmus, 2007; Wirtz et al., 2011). As such, this chapter adds to the **underresearched area of railway traffic controller fatigue**, while responding to the **fatigue risk modelling deficiencies** reported in the literature.

1.2.4 Chapter 4: Multi-output efficiency and human error

The final chapter presents a DEA-based benchmarking model, **tailored to computerized railway traffic control**, and **relates the obtained efficiency results with human error**. As such, with reference to the new research field of railway traffic control efficiency, initiated by the first article in this dissertation, we further contribute to the transportation literature.

The paper also contributes to the efficiency literature, in three distinct respects. First, by **calculating efficiency at an hourly resolution**, we assess the around-the-clock efficiency of work shift schedules, and reveal within-shift patterns of inefficiency. As such we offer an exceptionally disaggregated application of DEA. We have not identified any staffing efficiency study in the service sector examining efficiency levels at hourly, daily, or even weekly basis. Notable exceptions of temporally disaggregated efficiency studies can be however found in manufacturing (Hoopes and Triantis (2001) analyse at work order level; Jain et al. (2011)

at weekly level level) and in the fishing industry (e.g.,Vázquez-Rowe and Tyedmers (2013) analyse at weekly level). Our benchmarking model therefore provides the first application for evaluating hourly efficiency, not only for staffing efficiency but also in general. Providing quantitative insights in 'hour-of-day' and 'day-of-week' efficiency variations, the proposed approach empowers management to focus their attention to the most prominent staffing efficiency issues, and optimize their staffing levels and work shift patterns on an ex post basis. Aligning staff with workload in a more fair and equitable manner can not only improve efficiency, but also positively affect staff well-being and job satisfaction.

Second, to capture the traffic control process in a realistic way, we **customize and extend a multi-output efficiency measurement model (Cherchye et al., 2013, 2015)**, by formally including a priori information on the allocation of inputs to outputs. The Cherchye et al. (2013, 2015) multi-output methodology explicitly models the production technology (i.e., the set of feasible input-output combinations) for each individual output, while accounting for interdependencies between the different output-specific technologies. We formally include quantitative information on the production process in the model, by adding constraints of a proportional form, which have a straightforward, natural interpretation. Our method parallels the approach proposed by Cook and Zhu (2011), who recognized that the relative importance of inputs can be output-specific. They introduced output-specific constraints (weight restrictions) on the inputs, and encouraged further research along these lines. Importantly, and similar to the DEA virtual weight restrictions introduced by Wong and Beasley (1990), proportionally defined bounds are particularly appealing when a priori expert judgment needs to be translated into DEA restrictions (Sarrico and Dyson, 2004). Through a judicious choice of the upper and lower bounds for the input-output output allocations, we can easily and flexibly tailor the internal structure of the DMUs to the specific production process under evaluation. As such, by carefully adapting our methodology to address the traffic control problem at hand, we have conceived a more general framework, able to handle a wide range of real-life settings. Finally, by virtue of the applied multi-output methodology, we can further decompose the efficiency scores and pinpoint the operational reasons underlying the observed efficiency patterns.

The hourly and multi-output efficiency results reveal a 15 percentage point gap between

the (average) highest and lowest hourly efficiency levels. This suggests that management should evolve from a relatively inflexible and 'one-size-fits-all' scheduling philosophy, consisting of non-overlapping and fixed 8-hour shifts, to a more customized approach (e.g. by gradually changing team size and composition, revise shift length and starting times, and scheduling overlapping shifts). As such, efficiency gains are expected by **introducing a more flexible staff scheduling approach**, which can be actively supported by the efficiency estimations. Also, we find no evidence against the idea that **sufficient time is allocated to perform safety operations**.

Third, by linking the exceptionally disaggregated information on multi-output efficiency with equally disaggregated data on human error, our contribution to the literature extends **beyond the scope of traditional efficiency methods**. The observed (non safety-critical) task errors occur in 38% of the hourly observations. By estimating a series of Probit regression models, we pinpoint an empirical link between the components of multi-output efficiency and the probability of human error. Fully leveraging the insights provided by the multi-output cost efficiency framework, we reveal a relation with the output-specific efficiency of production tasks with a highly variable work load. For these production tasks, reaching the upper boundaries of the modelled input-output allocation restrictions also significantly impacts error probability: the binding allocation constraints are found to associate with a 3 to 4 percentage point higher probability of human error. The results therefore suggest that decision makers, when examining the **possibly detrimental effects of efficiency changes on safety levels**, should **focus on the staff allocations towards outputs with a highly variable and unpredictable workload**. This demonstrates the usefulness of multi-output efficiency models in providing insights that go beyond mere productive efficiency considerations.

In sum, the proposed approach and the exceptionally disaggregated data quantitatively support decision makers in focusing on key efficiency parameters, while keeping safety in the spotlight. This allows management to iteratively tackle efficiency issues and gradually move towards optimized hourly staffing levels. At the same time, it monitors safety operations and reveals relationships with human error. Finally, because of the large volumes of data involved, this chapter also aims to foster further research on developing DEA tools for large-

scale production data.

1.3 Data sources and Business Intelligence tool

Given the data-driven nature of the research, this dissertation was actively supported by a purpose-built Business Intelligence application. The application lays a **cornerstone of our research design**: the **link between workforce data and the corresponding production data**. The tool is directly connected to the (Belgian railway traffic control) work schedule and operational databases, links these data sources at different aggregation levels, prepares the necessary datasets for the performance analysis, and incorporates the intermediate and final results of the research.

As operational IT systems are generally not designed for ex-post performance analysis (Triantis, 2011), one of the heaviest challenges faced during the development of the supporting Business Intelligence tool was to translate the available data into useful measures of the traffic control process. In addition, as the workforce and operational systems were conceived with different objectives and users in mind (human resource management/scheduling versus engineering/operations), this required the construction of several additional data tables, linking the original data sources at the desired level of disaggregation.

Information technology has much evolved since Golany and Roll (1989) first suggested 'report generation' and 'graphical data analysis' in their influential 'DEA application procedure', and with the advent of Business Intelligence and Business Analytics software there are now a plethora of functions available for exploring and probing efficiency and fatigue risk results. Our highly interactive Business Intelligence tool fuelled the interaction with the railway experts, and as such **provided substantial leverage** for the iterative process of **model specification and face-validation**. In addition, in order to streamline and strengthen the interactions with the railway experts, particular attention was paid to the transparent translation of the proposed models in operationally intelligible language and metrics. This aspect was critical in terms of safeguarding the iterative character of the approach, which formed the backbone of the research design.

References

Åkerstedt, T. (2000). Consensus statement: fatigue and accidents in transport operations. *Journal of sleep research*, 9(4):395–395.

Andersson, M. (2008). Marginal railway infrastructure costs in a dynamic context. *European Journal of Transport and Infrastructure Research*, 4(8).

Banker, R. D. and Morey, R. C. (1986). The use of categorical variables in data envelopment analysis. *Management science*, 32(12):1613–1627.

Belenky, G., Wesensten, N. J., Thorne, D. R., Thomas, M. L., Sing, H. C., Redmond, D. P., Russo, M. B., and Balkin, T. J. (2003). Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: A sleep dose-response study. *Journal of sleep research*, 12(1):1–12.

Bilotkach, V., Gitto, S., Jovanović, R., Mueller, J., and Pels, E. (2015). Cost-efficiency benchmarking of european air navigation service providers. *Transportation Research Part A: Policy and Practice*, 77:50–60.

Brogmus, G. (2007). Day of the week lost time occupational injury trends in the us by gender and industry and their implications for work scheduling. *Ergonomics*, 50(3):446–474.

Button, K. and Neiva, R. (2014). Economic efficiency of european air traffic control systems. *Journal of Transport Economics and Policy*, 48(1):65–80.

Cabon, P., Deharvengt, S., Grau, J. Y., Maille, N., Berechet, I., and Mollard, R. (2012). Research and guidelines for implementing fatigue risk management systems for the french regional airlines. *Accident Analysis & Prevention*, 45:41–44.

Castelli, L., Pesenti, R., and Ukovich, W. (2010). A classification of DEA models when the internal structure of the decision making units is considered. *Annals of Operations Research*, 173(1):207–235.

Charnes, A., Cooper, W. W., and Rhodes, E. (1978). Measuring the efficiency of decision-making units. *European journal of operational research*, 2(6):429–444.

Cherchye, L., De Rock, B., and Walheer, B. (2015). Multi-output efficiency with good and bad outputs. *European Journal of Operational Research*, 240(3):872–881.

Cherchye, L., Rock, B. D., Dierynck, B., Roodhooft, F., and Sabbe, J. (2013). Opening the "black box" of efficiency measurement: input allocation in multioutput settings. *Operations Research*, 61(5):1148–1165.

Civity management consultants (2013). International benchmarking of network rail's operations and support functions expenditure. *Department for Transport and Office of Rail Regulation; London.*

Cooper, W. W., Seiford, L. M., and Zhu, J. (2011). *Data Envelopment Analysis: Handbook on Data Envelopment Analysis.* Springer.

Cotrim, T., Carvalhais, J., Neto, C., Teles, J., Noriega, P., and Rebelo, F. (2017). Determinants of sleepiness at work among railway control workers. *Applied ergonomics*, 58:293–300.

Cowie, J. and Loynes, S. (2012). An assessment of cost management regimes in british rail infrastructure provision. *Transportation*, 39(6):1281–1299.

Darwent, D., Dawson, D., Paterson, J. L., Roach, G. D., and Ferguson, S. A. (2015). Managing fatigue: It really is about sleep. *Accident Analysis & Prevention*, 82:20–26.

Dawson, D., Noy, Y. I., Härmä, M., Åkerstedt, T., and Belenky, G. (2011). Modelling fatigue and the use of fatigue models in work settings. *Accident Analysis & Prevention*, 43(2):549–564.

Dean, D. A., Fletcher, A., Hursh, S. R., and Klerman, E. B. (2007). Developing mathematical models of neurobehavioral performance for the "real world". *Journal of Biological Rhythms*, 22(3):246–258.

Di Milia, L., Smolensky, M. H., Costa, G., Howarth, H. D., Ohayon, M. M., and Philip, P. (2011). Demographic factors, fatigue, and driving accidents: An examination of the published literature. *Accident Analysis & Prevention*, 43(2):516–532.

Dorrian, J., Baulk, S. D., and Dawson, D. (2011). Work hours, workload, sleep and fatigue in australian rail industry employees. *Applied ergonomics*, 42(2):202–209.

Fletcher, A. and Dawson, D. (2001). Field-based validations of a work-related fatigue model based on hours of work. *Transportation research part F: traffic psychology and behaviour*, 4(1):75–88.

Folkard, S. and Lombardi, D. A. (2004). Toward a "risk index" to assess work schedules. *Chronobiology international*, 21(6):1063–1072.

Folkard, S. and Lombardi, D. A. (2006). Modeling the impact of the components of long work hours on injuries and "accidents". *American journal of industrial medicine*, 49(11):953–963.

Folkard, S., Robertson, K. A., and Spencer, M. B. (2007). A fatigue/risk index to assess work schedules. *Somnologie-Schlafforschung und Schlafmedizin*, 11(3):177–185.

Fried, H. O., Lovell, C. K., and Schmidt, S. S. (2008). *The measurement of productive efficiency and productivity growth*. Oxford University Press.

Gander, P., Hartley, L., Powell, D., Cabon, P., Hitchcock, E., Mills, A., and Popkin, S. (2011). Fatigue risk management: Organizational factors at the regulatory and industry/company level. *Accident Analysis & Prevention*, 43(2):573–590.

Gertler, J., DiFiore, A., and Raslear, T. (2013). Fatigue status of the us railroad industry. Technical Report DOT/FRA/ORD-13/06, US Department of Transportation, Federal Railroad Administration, Washington DC.

Golany, B. and Roll, Y. (1989). An application procedure for DEA. *Omega*, 17(3):237–250.

Greubel, J. and Nachreiner, F. (2013). The validity of the risk index for comparing the accident risk associated with different work schedules. *Accident Analysis & Prevention*, 50:1090–1095.

Hoopes, B. J. and Triantis, K. P. (2001). Efficiency performance, control charts, and process improvement: complementary measurement and evaluation. *Engineering Management, IEEE Transactions on*, 48(2):239–253.

International Union of Railways (2002). Infracost-the cost of railway infrastructure. *Final Report, Paris*.

Jain, S., Triantis, K. P., and Liu, S. (2011). Manufacturing performance measurement and target setting: A data envelopment analysis approach. *European Journal of Operational Research*, 214(3):616–626.

Johansson, P. and Nilsson, J.-E. (2004). An economic analysis of track maintenance costs. *Transport Policy*, 11(3):277–286.

Kennedy, J. and Smith, A. S. (2004). Assessing the efficient cost of sustaining britain's rail network: Perspectives based on zonal comparisons. *Journal of Transport Economics and Policy*, 38(2):157–190.

Kneip, A., Simar, L., and Wilson, P. W. (2008). Asymptotics and consistent bootstraps for DEA estimators in nonparametric frontier models. *Econometric Theory*, 24(06):1663–1697.

Monk, T. H. and Wagner, J. A. (1989). Social factors can outweigh biological ones in determining night shift safety. *Human Factors*.

Popkin, S., Gertler, J., and Reinach, S. (2001). A preliminary examination of railroad dispatcher workload, stress, and fatigue. Technical report.

Rail Safety and Standards Board (2015). Fatigue and its contribution to railway incidents.

Sarrico, C. S. and Dyson, R. (2004). Restricting virtual weights in data envelopment analysis. *European Journal of Operational Research*, 159(1):17–34.

Simar, L. and Wilson, P. W. (2007). Estimation and inference in two-stage, semi-parametric models of production processes. *Journal of econometrics*, 136(1):31–64.

Smith, A., Wheat, P., and Smith, G. (2010). The role of international benchmarking in developing rail infrastructure efficiency estimates. *Utilities policy*, 18(2):86–93.

Smith, A. S. (2012). The application of stochastic frontier panel models in economic regulation: Experience from the european rail sector. *Transportation Research Part E: Logistics and Transportation Review*, 48(2):503–515.

Triantis, K. P. (2011). Engineering applications of data envelopment analysis. In *Handbook on Data Envelopment Analysis*, pages 363–402. Springer.

Tucker, P., Folkard, S., et al. (2012). *Working time, health and safety: A research synthesis paper*. International Labour Organization.

Vázquez-Rowe, I. and Tyedmers, P. (2013). Identifying the importance of the "skipper effect" within sources of measured inefficiency in fisheries through data envelopment analysis (DEA). *Marine Policy*, 38:387–396.

Wheat, P. and Smith, A. S. (2008). Assessing the marginal infrastructure maintenance wear and tear costs for britain's railway network. *Journal of Transport Economics and Policy*, 42(2):189–224.

Williamson, A., Lombardi, D. A., Folkard, S., Stutts, J., Courtney, T. K., and Connor, J. L. (2011). The link between fatigue and safety. *Accident Analysis & Prevention*, 43(2):498–515.

Wilson, J. R. and Norris, B. J. (2005). Rail human factors: Past, present and future. *Applied ergonomics*, 36(6):649–660.

Wirtz, A., Nachreiner, F., and Rolfes, K. (2011). Working on sundays–effects on safety, health, and work-life balance. *Chronobiology international*, 28(4):361–370.

Wong, Y.-H. and Beasley, J. (1990). Restricting weight flexibility in data envelopment analysis. *Journal of the Operational Research Society*, pages 829–835.

Chapter 2

RAILWAY TRAFFIC CONTROL EFFICIENCY AND ITS DETERMINANTS

Abstract

The present paper fills a gap in the literature by examining the efficiency of railway traffic control. In spite of large-scale migration strategies towards centralised signal boxes (traffic control centres), railway traffic control still remains a labour-intensive process in many European countries. In close collaboration with experts from Infrabel, the Belgian railway infrastructure manager, we develop a two-stage benchmarking framework which assesses and explains railway traffic control efficiency. In the first stage, a bootstrapped Data Envelopment Analysis model with categorical variable assesses efficiency, and closely monitors average and individual performance trends over time. Second-stage regressions examine the impact of several factors on efficiency. The proposed framework can be adopted by infrastructure managers as an internal benchmarking tool, evaluating the entire network or specific sub-regions. We demonstrate the practical applicability of our approach with a unique and rich 18-month dataset of Infrabel's relay-technology signal boxes. Aiming to uncover additional insights, we perform our analysis on two subsets of the monthly data: one covering the working week, the other the weekends. Our findings suggest that in order to improve on traffic control efficiency, railway infrastructure managers should aim for geographical concentration, larger team sizes, and a continuous follow-up of signal box opening times. Further efficiency gains can be generated by reducing infrastructure complexity. Finally, our results also indicate that railway infrastructure managers should take into account the differences between working week and

The work in this chapter is co-authored by Johan Christiaens.

weekend when measuring and analysing traffic control performance.

Keywords:bootstrap; Data Envelopment Analysis; efficiency; railway infrastructure; traffic control; two-stage approach

2.1 Introduction

Railway infrastructure managers are increasingly urged by European railway directives and national austerity measures to improve on their efficiency levels. Clearly illustrating this, the European Directive 2012/34/EU (2012)[1] on the establishment of a Single European Railway Area states that 'railway infrastructure is a natural monopoly and it is therefore necessary to provide infrastructure managers with incentives to reduce costs and to manage their infrastructure efficiently.' The same directive also defines an infrastructure manager as 'any body or firm responsible in particular for establishing, managing and maintaining railway infrastructure, including traffic management and control-command and signalling'.

Scholarly research on the efficiency of what is typically referred to as railway infrastructure asset management, i.e. establishing, managing and maintaining the infrastructure, was initiated with the internal benchmark of Network Rail's[2] maintenance and renewal zones by Kennedy and Smith (2004). Their analysis was followed by a series of international studies, all focusing on asset management efficiency (see e.g. Smith et al., 2010). Railway traffic control, however, consistently remained out of scope of all previous research. The present paper addresses this void in the literature. In support of the research, Ghent University initiated a research project, baptised CRIPTON[3], together with Belgian railway infrastructure manager Infrabel.

For the purpose of this paper, we define railway traffic control as the combination of real-time traffic management (i.e. real-time decision making by dispatchers to ensure a fluent traffic flow) and signalling activities (i.e. the authorisation of train movements through the signalling system, by signallers). Although of high importance, the technical and engineering

[1]More commonly known as the 'recast' of the first railway package.

[2]The British railway infrastructure manager.

[3]CRIPTON = Comprehensive Railway Infrastructure Productivity Tools for Operations on the Network.

aspects of the systems supporting the traffic control activities, i.e. the train control-command and signalling systems, are not the subject of this research. The need to provide an own definition of railway traffic control stems from the diversity of systems and procedures across Europe, and the corresponding disparity in terminology (Pachl, 2009, preface). For the remainder of this paper, we will adhere as closely as possible to the glossary on railway operation and control developed by (Pachl, 2009).

Railway traffic control is performed at several levels, ranging from central to local. Our research is focusing on the traffic control activities performed in the so-called interlocking stations or signal boxes. Railway staff working in these signal boxes are mainly responsible for local or regional traffic management and signalling. At present, several European infrastructure managers are migrating the technology behind these signal boxes from the existing legacy systems (mechanical, electro-mechanical, relay-based or other technologies) towards a more modern and computerised environment, in which centralisation and automation are the keywords.

Reliable information on these migration projects, as well as their current status, is rather fragmented. A relatively good overview is provided in the recent benchmarking report published by the UK Office of Rail Regulation(Civity management consultants, 2013). With 7 infrastructure managers participating in the study, the report states that the levels of traffic control centralisation and automation vary significantly across Europe. It identifies a series of leaders with a high degree of centralisation (such as the infrastructure manager ProRail in the Netherlands) and followers, fully progressing in ambitious modernisation projects (e.g. Network Rail in the UK, RFF[4] in France, or Infrabel in Belgium). It is clear however that, despite these large-scale and long term investment projects, railway traffic control currently still remains a labour-intensive process in many European countries. For instance, Network Rail aims to replace its 800 signal boxes by 14 Rail Operating Centres in a migration project stretching over several decades. The French infrastructure manager RFF targets the year 2030 to centralise their 1.500 signal boxes and 21 regional centres in 16 traffic control centres. In Belgium, Infrabel strives to replace its legacy system signal boxes in 10 electronic signal boxes by 2022. Infrabel staff involved in real-time traffic control currently adds up to about

[4]Réseau Ferré de France, merged into SNCF Réseau as of 1 January 2015.

1800 persons, spread over up to 120 signal boxes (average number in 2013)[5].

The main contributions of this paper are threefold. First, drawing on related research as well as railway expertise, we present a benchmarking framework based on Data Envelopment Analysis (DEA), which assesses and closely monitors the efficiency of traffic control in signal boxes. The second-stage regressions of the framework examine the impact of several environmental factors on efficiency. The framework can be adopted by other infrastructure managers as an internal benchmarking tool, evaluating the entire network or specific subregions. Second, we demonstrate the practical applicability of our framework with a unique and rich 18-month dataset of relay-technology signal boxes provided by Infrabel. Aiming for additional insights, we perform our analysis on two subsets of the monthly data: one covering the working week, the other the weekends. Third, not only do our empirical findings suggest the significant impact of a number of environmental factors on efficiency, they also show differences between working week and weekend efficiency. The results are expected to be generalizable to other signal box technologies, and to railway networks or regions with a comparable range of traffic density and infrastructure complexity.

The remainder of this paper is structured as follows. Section 2 provides an overview of related research. In the methodology section we then model the traffic control production process, as well as the environmental variables influencing its efficiency, and present the DEA-based two stage approach. Section 4 describes the data for the practical application of the benchmarking framework, and section 5 reports and discusses the empirical results. Conclusions and recommendations for railway infrastructure managers are set out in the final section.

2.2 Related research

The gradual vertical separation of infrastructure and train operations, one of the cornerstones of Europe's railway policy, has increased the academic attention towards the cost and efficiency of railway infrastructure. The existing body of literature in this research area has been steadily complemented by specific infrastructure oriented research, with a main focus on marginal cost estimation (e.g. Johansson and Nilsson, 2004; Wheat and Smith, 2008;

[5]Source: Infrabel data.

Andersson, 2008) and efficiency measurement (e.g. Kennedy and Smith, 2004; Smith, 2012; Smith and Wheat, 2012). The scope of this previous research on the cost and efficiency of railway infrastructure was limited to asset management, and was almost exclusively based on parametric techniques.

Although the subject of railway traffic control is gradually emerging in scholarly publications, research on its efficiency has not yet been performed. Railway traffic control does appear in fragments in previous research, but always within the context of a broader research topic (such as the impact of vertical separation on efficiency), and is referred to under a variety of terms. Clearly, the disparity in terminology across Europe is also reflected in scholarly research.

For instance, in an efficiency analysis of European railways, Growitsch and Wetzel (2009) apply a bootstrapped Data Envelopment Analysis model to examine the economies of scope associated with the vertical separation of infrastructure management and train operations. One of the theoretical elements cited, is the cost of 'real-time traffic coordination'. Research by Merkert and Nash (2013) investigates on the size and nature of transaction costs between infrastructure managers and train operators (a consequence of the vertical separation). Based on in-depth interviews with senior rail managers, the study calls attention to 'control centres of the infrastructure manager' and 'real-time decisions' as elements in the complex and intense area of 'day-to-day operations'. A paper by Cowie and Loynes (2012), analysing the evolution of British railway infrastructure costs over the years 1980-2009, mentions 'controlling traffic movements' through operation of the signalling system as 'the second component of operating the infrastructure'. As a final example, Hansen et al. (2013) discuss a series of Key Performance Indicators relevant for international benchmarking of train operations as well as infrastructure management. The authors suggest the further breakdown of infrastructure management activities and costs into general administration, maintenance and repair, 'traffic control' and investment projects.

Only at an industry level, a slowly increasing number of reports explicitly examining railway traffic control efficiency are emerging. First in line of a series of studies was a chapter on 'operation management cost' in the InfraCost study (International Union of Railways, 2002), in which the initial steps towards a benchmarking methodology were taken. For the 14

Western European companies participating in the project, the yearly operation management cost rose to an 8-9 billion EUR order of magnitude, which represented about 30 % of total annual expenditures for infrastructure management (based on year 2000 budgets). Labour cost was the dominating factor in operation management and represented, on average, about 90 %. Based on additional data gathered from a more restricted sample of 10 UIC members, a number of partial productivity ratios (such as operation management cost per maintrack-km or per train-km) were presented in an anonymized reporting.

And finally, similar benchmarking work was carried out by the same group of consultants[6] within the context of the McNulty (2011) Value For Money report, and in a further extension of this study in a benchmarking report for the UK Office of Rail Regulation (Civity management consultants, 2013). The latter report advocates that optimal migration strategies for railway traffic control should consist of a combination of both centralisation and, in parallel, optimisation of staffing levels. A series of measures to achieve this are put forward, e.g. more sophisticated staffing calculations, part-time work, and the optimal alignment of rostering plans to the traffic profile.

Most probably the major cause for the current neglect of railway traffic control efficiency in the scholarly literature is the lack of sufficiently disaggregated - or even basic - data. In the area of air traffic control research, much data is publicly available through the annual benchmarking reports provided by the EUROCONTROL Performance Review Commission (ATM Cost-Effectiveness Benchmarking Reports). In addition, EUROCONTROL has commissioned a series of parametric and non-parametric studies in the past years to assess the efficiency of Air Navigation Service Providers (e.g. Mouchart and Simar, 2002; Holder et al., 2006; EUROCONTROL Performance Review Commission, 2011). Recently, two academic articles have been published (Button and Neiva, 2013, 2014), benchmarking the European Air Navigation Service Providers against each other by means of bootstrapped DEA, and analysing the environmental variables influencing efficiency in a second stage regression.

Although the model specifications, results and conclusions of the air traffic control research cannot be directly transposed our research area, they provide a valuable source of information for our benchmarking framework. For an overview of the main similarities and differences

[6]BSL, and later on civity Management Consultants.

between air and rail traffic control, we refer to Pellegrini and Rodriguez (2013).

2.3 Methodology

In the first stage of our benchmarking framework, we estimate traffic control efficiency by means of Data Envelopment Analysis (DEA). The DEA methodology is a powerful non-parametric tool for assessing the efficiency of operational processes with multiple inputs and outputs (Cooper et al., 2011). In the second stage of the framework, we apply second-stage regressions to examine the impact of environmental variables on the obtained efficiency scores. Environmental variables are factors that could influence efficiency, but are assumed not under the control of management (Coelli et al., 2005). This two-stage approach allows for hypothesis tests on the effects of the environmental variables, and can be considered as more transparent than the alternative, i.e. including these variables in the DEA model specification (ibid.).

As one of the objectives of the benchmarking framework is to keep close track of traffic control performance, the developed model is based on a monthly evaluation of efficiency, but can easily be adapted to other monitoring frequencies. In addition, in order to assure a fair efficiency comparison, only signal boxes equipped with the same technology (e.g. electro-mechanical, relay-based, or electronic) should be benchmarked against each other.

In order to support the model building process, an expert panel composed of Infrabel specialists from operations, accounting and data departments was established (see Golany and Roll (1989) for a DEA application procedure invoking expert knowledge). The panel provided valuable feedback on previous related research and its applicability on railway traffic control. Moreover, much attention was paid to the understanding of the DEA concept by the experts. The intuition behind the methodology was carefully explained and visualised, without diving into the mathematical details. This was a critical step in interpreting the results and acknowledging potential limitations of the analysis (Ozbek et al., 2009).

In the remainder of this section we will model the traffic control production process, through a definition of its inputs and outputs. Also, we will present the environmental variables expected to influence its efficiency, and discuss the decision-making levels related to these variables. Finally, the DEA-based two-stage methodology will be detailed.

2.3.1 Model specification

The traffic control production process

To define the traffic control production process in the signal boxes, we specify a model with one input and multiple outputs. The hours worked in the signal boxes serve as the single input, while the output mix consists of two types of services: two outputs capture the workload associated with railway traffic (train and shunting movements), while two other variables account for the workload related to the railway infrastructure (lines and nodes of the network).

The local management of the signal boxes has no control over the exogenously determined traffic and infrastructure outputs but it holds, within the limits of its own authority, responsibility for the optimal alignment of the inputs (i.e. the hours worked by signal box staff) with these outputs. As the signal boxes are benchmarked against others equipped with the same technology, we do not consider other inputs such as technical properties or capital expenditures[7]. At a central decision-making level, senior management responsible for traffic control policy can apply the developed benchmarking model to not only capture best and worst traffic control practices across their network, but also to closely monitor the evolution of the efficiency scores. We will now proceed with a detailed description of the input and output variables of the production process.

HOURS. This single input fully captures the resources lined up for signalling and traffic management, and is defined as the total number of hours worked in the signal box, by the dispatchers and signallers. Their tasks also include monitoring the infrastructure, safety measures in case of infrastructure works, and the attribution of delay causes to the infrastructure manager or the train operators. There is no outsourcing involved, neither in the Infrabel case, nor in any other European case known to the authors and the Infrabel experts. Sometimes both functions of signalling and traffic management are performed by the same person. The so-called available time, which reflects the free time between tasks (e.g. in signalling), is included in this variable.

[7]This approach is in line with all the above-mentioned international studies from the railway sector, in which the operational expenditures (predominantly labour costs) are benchmarked, see International Union of Railways, 2002; McNulty, 2011; Civity management consultants, 2013.

TRAIN and SHUNT. The first two outputs account for the workload associated with movements of railway vehicles. These movements can be divided in train and shunting movements (Pachl, 2009). Shunting involves all movements other than train movements (e.g. train formation, shunting from sidings to station tracks and back), is performed at low speed, and follows operating procedures different from train movements. The first output TRAIN accounts for the signalling, traffic management and delay attribution of the train movements. The variable counts these movements in each network node (i.e. station or junction), and is weighted according to the corresponding workload for the signal box: trains passing through nodes without any stop have a weight of 1; trains with arrival and departure receive a weight of 2, as this requires two separate route settings and dispatching efforts for the train. The SHUNT variable accounts for the shunting workload in the signal boxes. A concern in modelling this workload may be the absence of data to capture the shunting movements[8]. To circumvent this issue, we define the SHUNT variable as an ordered categorical variable. Together with the expert panel, we determined 5 levels of shunting workload, relative to the total number of train and shunting movements. The highest shunting workload (level 1 of the variable) is attributed to signal boxes in which shunting represents 100% - 80% of total movements, level 2 accounts for a shunting workload of 80% - 60%, and so on until level 5, in which shunting is assessed as representing 20% - 0% of total movements.

LINES and NODES. While the first outputs are related to the active role of signallers and dispatchers, a second category of variables captures the more passive character of the activities in the signal boxes. Surveillance of infrastructure components such as switches, signals, level crossings, track circuits used for train detection, as well as tasks related to ensuring the safety of infrastructure works are brought into the model through the LINES and NODES variables. They are defined as the number of main line kilometres (LINES) and the number of stations and junctions (NODES) controlled by the signal box. In order to ensure correct comparability across months, these variables were added up on a daily basis. In addition, as signal boxes can be closed for a period of time, a fair and equitable benchmarking also requires each daily line length and number of nodes to be multiplied with the percentage of time the signal box

[8]As shunting movements are executed inside stations, sufficiently detailed data on the shunting movements authorised by the signal boxes may not always be available to the infrastructure manager (with the exception of electronic signal boxes).

is in operation. For example, if a signal box is open for 80% of a certain day, the monitored lines and nodes can only be considered an output for the same percentage of this day, and are correspondingly multiplied by 0.80. As we shall see in the discussion on the environmental variables, the opening and closing of signal boxes (and hence of the infrastructure) is set by central management. Within these exogenously determined time limits, local management has the responsibility of optimally aligning the HOURS input (i.e. the sum of all hours worked by the staff in the signal box) with the traffic and infrastructure outputs. Ideally, both the LINES and NODES variables should also capture the partial closing of the network (within the area controlled by the signal boxes). Similar to the opening or closing of airspace sectors in Air Navigation Service Providers, signal boxes can be closed when traffic volumes do not justify the presence of staff (see also International Union of Railways (2002)).

Environmental variables

The factors expected to influence railway traffic control efficiency can be grouped in 3 categories of environmental variables. The first group represents traffic and timetable characteristics (e.g. traffic density), and is considered exogenous to infrastructure manager. The other environmental variables are related to the infrastructure manager's internal decision-making, and can be subdivided into two distinct categories. First, we have identified variables corresponding with the asset management component of railway infrastructure (e.g. track layout complexity). Second, we consider variables which reflect decisions made by the central management responsible for traffic control (e.g. signal box closing times). All environmental variables are beyond the control of local management of the signal boxes, but are expected to influence efficiency levels in a positive or negative way. We will now describe these variables one by one.

The first group of environmental variables can be considered as being exogenous to the infrastructure management, and contains traffic and timetable characteristics. These variables are largely influenced by macro-economic factors or public service requirements (see e.g. Merkert et al. (2010), and the corresponding timetable put forward by the railway undertakings. VAR is accounting for the variability of the hourly traffic profile, a factor expected to have a negative impact on efficiency levels. In accordance with the EUROCONTROL econometric models (2011), we calculate the variability by dividing the maximum number of

weighted train movements per hour (i.e. during the hourly peak) by the average number per hour (during opening times). Traffic density is introduced through two variables DENS_SPAT and DENS_TEMP, respectively reflecting spatial and temporal density of traffic. Spatial density (DENS_SPAT) is expected to increase efficiency, as it reduces the amount of available time in the signal boxes, while higher temporal densities (DENS_TEMP) are expected to exert a negative influence. DENS_SPAT is calculated as the number of weighted train movements TRAIN divided by the NODES variable. We proxy the temporal traffic density DENS_TEMP by the number of secondary delays on the line (i.e. delays passed from one train to another), divided by the number of weighted train movements TRAIN. We also examine the impact of several timetable properties (i.e. train connections and changes in rolling stock or train crew, performed at the station platforms), through the TT.CHAR variable. These characteristics complexify the decisions to be taken in the signal box, as well as their timely execution, and are therefore expected to decrease efficiency. We proxy the TT_CHAR variable through the number of train delays due to these connections or changes, divided by the number of weighted train movements TRAIN.

The next category of variables examines the impact of asset management policy on traffic control efficiency. First, the reduction of infrastructure complexity was put forward as an important lever for improving traffic control efficiency, not only by the Infrabel experts, but in also previous work (International Union of Railways, 2002). In our study, we consider two levels of complexity: a higher level COMP_NET, reflecting the complexity of the railway network under the control of the signal box, and a second level COMP_TRACK, capturing the complexity of the track layout. The COMP_NET variable is calculated as the number of nodes divided by the number of lines, while COMP_TRACK is proxied by the number of signals divided by the number of nodes. Intermediate block signals (automatic signals between signal boxes) and dwarf signals (small ground mounted signals located at sidings) are not taken into account, as they do not add to the complexity of the train movements and could bias the calculation of the complexity ratio. Second, we also examine the proportion of stations in the network, relative to the number of nodes (i.e. stations and junctions). The variable P_STATIONS is expected to exert a negative effect on efficiency, due to the additional complexity in handling train movements. A final variable linked to the infrastructure, WORK_DENS, represents the density of infrastructure works (i.e. maintenance and renewal

of tracks, switches, catenaries, and signalling equipment). We proxy this variable through the number of delays caused by infrastructure works, divided by the length of the lines during opening times (LINES variable).

The final group of environmental variables captures decisions made by the central management responsible for traffic control. First, as reducing opening times of the signal box is expected to increase efficiency levels, we introduce the P_CLOSED variable. It is defined as the percentage of signal box closing time relative to the total considered time (e.g. all days of the month x 24 hours). The opening and closing of infrastructure (lines and stations) through the opening and closing of signal boxes is a decision taken at central level, and can affect several signal boxes along the concerned railway axes. Second, the team size in the signal box is expected to increase efficiency, as the alignment of staffing levels in the signal box with the hourly traffic profile can be easier attained in larger signal boxes (Civity management consultants, 2013). We proxy team size by N_PERSONS, the total number of traffic controllers (dispatchers, signallers) who worked in the signal box during the month under consideration. Third, verifying the impact of geographical centralisation, the KM_PERSON variable (LINES divided by N_PERSONS) is expected to display a positive sign in the regression results. The last two variables assumed to be largely controllable by central management check for a possible association between human factors and efficiency, without a priori expectations on the sign of the possible effects. AVG_AGE is the average age of the staff who worked in the signal box, and serves as a proxy for their experience and skills. The ERRORS variable captures the human errors by signal box staff leading to quality issues. It is calculated as the number of delays due to these human errors, divided by the number of weighted movements TRAIN, and rescaled upwards by a factor of 1000.

2.3.2 Data Envelopment Analysis model with categorical variable

In the first stage of our analysis, we estimate the relative efficiency of the signal boxes by means of DEA. The DEA methodology can briefly be described as 'a data-oriented approach for evaluating the performance of a set of peer entities called Decision-Making Units (DMU), which convert multiple inputs to multiple outputs' (Cooper et al., 2011). Applying mathematical programming techniques, DEA evaluates the relative efficiency of these DMU (the signal boxes in our analysis) with a minimum of a priori assumptions. These assumptions are

generally referred to as the free disposability (i.e. the possibility of producing less outputs with more inputs) and convexity of the examined technology (i.e. a convex linear combination of the observed input-output combinations is also feasible). Based on these assumptions, DEA constructs an empirical production set $\widehat{\Psi}$, which contains all observed input-output combinations and which estimates the true attainable production set Ψ (i.e. the set of all physically attainable input-output combinations). The so-called technical efficiency of a specific DMU is then estimated relative to the boundary or production frontier $\delta\widehat{\Psi}$ of $\widehat{\Psi}$ (Simar and Wilson, 2008). For a more detailed discussion on DEA, we refer to the cited references.

In our analysis, as we expect scale to play a role in shaping the true production set Ψ, and as local management does not have the power to change the size of the signal boxes, we will apply the DEA Variable Returns to Scale (VRS) model. This model, introduced by Banker et al. (1984), takes scale differences into account when determining the production frontier $\delta\widehat{\Psi}$, and assures that a DMU is benchmarked against DMUs of the same scale. Also, as the local management is accountable for the optimal alignment of the inputs with the traffic and infrastructure outputs (which are uncontrollable by local management), we adopt an input-orientation to measure technical efficiency. I.e., the distance of a DMU to the empirical production frontier $\delta\widehat{\Psi}$ is determined by moving towards this frontier through contraction of its inputs (hours worked), while keeping the outputs at the same levels.

Also, given the objective of monitoring efficiency on a monthly basis, we consider each monthly observation to be a distinct DMU for the DEA calculations. In doing so, the efficiency of each DMU is gauged against a single empirical frontier, spanning all observations. Under the assumption of no technological change, this intertemporal approach (Tulkens and Vanden Eeckaut, 2006) presents the advantage of comparing each signal box not only with others but also against itself over time, allowing for additional insight in seasonal effects and trends (Boussofiane et al., 1991).

To incorporate the ordered categorical variable SHUNT (a variable capturing the shunting output), we apply the DEA model with categorical non-discretionary variables. This model was introduced by (Banker and Morey, 1986), as an extension to the basic DEA models. The categorical variable can assume one of L levels (1, 2, ..., L), which reflect the different conditions in which the DMU have to operate. A higher level refers to a more advantageous

environment. Each DMU is then evaluated against the empirical production frontier which envelopes its own category and all preceding (lower) categories. Thus, resting on the assumption that there is a natural nesting or hierarchy of the L categories, each unit is only compared with DMU operating under the same or harsher conditions (Cooper et al., 2011). The (Banker and Morey, 1986) model can also be applied in cases where not the environment, but one of the production inputs or outputs is a categorical variable. For instance, when the research output for universities is assessed in terms of 'good', 'better' or 'excellent' (see Boussofiane et al. (1991), or when the output 'quality of service' of municipalities is classified as 'good', 'normal' or 'bad' (Teresa Balaguer-Coll and Prior, 2009).

Several approaches are available for integrating the categorical variable in the DEA models (Löber and Staat, 2010). As there is only one categorical variable in our model, we simply apply different VRS frontiers for each level of the variable. Formally, based on the notations in Cooper et al. (2011), we calculate the efficiency estimate $\hat{\theta}^{l}_{VRS}(\mathbf{x}, \mathbf{y})$ for a DMU in level l of the categorical variable, and with input and output vectors $\mathbf{x}$ and $\mathbf{y}$, by solving the following linear programming model:

$$\hat{\theta}^{l}_{VRS}(\mathbf{x}, \mathbf{y}) =$$
$$min\left\{\theta > 0 \mid \theta\mathbf{x} \geqslant \sum_{i \in \bigcup_{f=1}^{l} K_f} \lambda_i \mathbf{x}_i; \mathbf{y} \leqslant \sum_{i \in \bigcup_{f=1}^{l} K_f} \lambda_i \mathbf{y}_i; \sum_{i \in \bigcup_{f=1}^{l} K_f} \lambda_i = 1; \lambda_i \geqslant 0; i = 1, ..., n\right\}, \tag{2.1}$$

where the sample of n observations $K \in \{1, 2, \ldots, n\}$ is split into L subsets $K_f = \{j \mid j \in K$ and level of the categorical variable $= f\}$, and $K_i \bigcap K_j = \varnothing, i \neq j$. In this equation, $(\mathbf{x}_i, \mathbf{y}_i)$ are the input and output vectors of the n observations in the sample. The scalars λ_i are the weights applied in the optimization problem to construct the empirical frontier $\delta\widehat{\Psi}^{l}_{VRS}$ which, under the assumption of Variable Returns to Scale, tightly envelops all observations of level l and lower.

2.3.3 Bootstrapping the efficiency estimates

As the efficiency scores are based on an empirical frontier, and not on the unknown true production frontier, these estimations are upward biased by construction: the probability of

including truly efficient units in the sample decreases with diminishing sample size, shifting the empirical frontier away from the true frontier. In addition, DEA efficiency scores are serially correlated in an unknown and complex way (Simar and Wilson, 2007). To deal with these issues, and before engaging in the second stage regression analysis, we apply the subsample bootstrapping algorithm proposed in the literature to obtain bias-corrected efficiency estimates (see Simar and Wilson (2008) for an overview and further technical details on this subject).

Now coming back to our traffic control model, the integration of the categorical variable (SHUNT) implies that the convexity assumption is relaxed for this dimension of the DEA model (Banker and Morey, 1986). We therefore need to adapt the bootstrap algorithm to accommodate for the l levels of the categorical variable. We do this by sampling in a way similar to the group-wise subsampling approach developed by Simar and Zelenyuk (2007). The details of the adapted algorithm are presented in the remainder of this section, and can be skipped by readers with no specific interest in its technicalities.

Let us first consider the DEA Variable Returns to Scale (VRS) model, introduced by Banker et al. (1984). In the input-oriented case, an estimate $\hat{\theta}_{VRS}(\mathbf{x}, \mathbf{y})$ of the true efficiency $\theta_{VRS}(\mathbf{x}, \mathbf{y})$ can be calculated by solving the following linear programming model:

$$\begin{aligned} &\hat{\theta}_{VRS}(\mathbf{x}, \mathbf{y}) = \\ &min\left\{ \theta > 0 \mid \theta \mathbf{x} \geqslant \sum_{i=1}^{n} \lambda_i \mathbf{x}_i; \mathbf{y} \leqslant \sum_{i=1}^{n} \lambda_i \mathbf{y}_i; \sum_{i=1}^{n} \lambda_i = 1; \lambda_i \geqslant 0; i = 1, ..., n \right\}. \end{aligned} \tag{2.2}$$

Here, $(\mathbf{x}_i, \mathbf{y}_i)$ are the input and output vectors of the n observations in the sample, and the scalars λ_i are the weights applied in the optimization problem (2.2) to construct the convex and free-disposal hull $\widehat{\Psi}$, tightly enveloping these observations.

The idea behind the bootstrap procedures is to approximate the unknown sampling distribution of $\hat{\theta}_{VRS}(\mathbf{x}, \mathbf{y}) - \theta(\mathbf{x}, \mathbf{y})$ through the empirical distribution of $\hat{\theta}^*_{VRS}(\mathbf{x}, \mathbf{y}) - \hat{\theta}_{VRS}(\mathbf{x}, \mathbf{y})$, in which $\hat{\theta}^*_{VRS}(\mathbf{x}, \mathbf{y})$ represents pseudo efficiency scores generated by the bootstrapping algorithm.

The standard naive bootstrap, where a set S^*_n of n pseudo-observations is randomly drawn

(independently, uniformly, and with replacement) from the original set of observations S_n and is subsequently used to calculate $\hat{\theta}^*_{VRS}(\mathbf{x}, \mathbf{y})$, is known to be inconsistent[9] . Two solutions providing consistent inference have been proposed by Kneip et al. (2008): a subsampling procedure and a smoothing technique. Of the two, the subsampling approach is the least complex to implement and allows for speedier computations, since it only differs from the naive bootstrap in the size of the pseudo-samples, by drawing $m < n$ instead of n pseudo-observations from S_n.

In each iteration b of the subsample bootstrap algorithm, the efficiency score $\hat{\theta}^*_{VRS,m,b}(\mathbf{x}, \mathbf{y})$ is calculated with the bootstrap sample $S^*_{m,b} = \left\{(x^{*,b}_i, y^{*,b}_i), i = 1, \ldots, m\right\}$ determining the bootstrap production possibility set:

$$\hat{\theta}^*_{VRS,m,b}(\mathbf{x}, \mathbf{y}) = \\ min\left\{\theta > 0 \mid \theta\mathbf{x} \geqslant \sum_{i=1}^{m} \lambda_i \mathbf{x}^{*,b}_i; \mathbf{y} \leqslant \sum_{i=1}^{m} \lambda_i \mathbf{y}^{*,b}_i; \sum_{i=1}^{m} \lambda_i = 1; \lambda_i \geqslant 0; i = 1, ..., m\right\}. \tag{2.3}$$

For the subsample bootstrap, Kneip et al. (2008) have proven that as the number of bootstrap iterations $B \to \infty$, the Monte Carlo empirical distribution of $m^{\frac{2}{N+M+1}}\Big(\hat{\theta}^*_{VRS}(\mathbf{x}, \mathbf{y}) - \hat{\theta}_{VRS}(\mathbf{x}, \mathbf{y})\Big)$ approximates the exact but unknown sampling distribution of $n^{\frac{2}{N+M+1}}\Big(\hat{\theta}_{VRS}(\mathbf{x}, \mathbf{y}) - \theta(\mathbf{x}, \mathbf{y})\Big)$. This given S_n , and in- and output dimensions N and M of the DEA VRS model, see Simar and Wilson (2008).

Turning now to the DEA model with a categorical variable (see equation 2.1), the integration of the variable implies that the convexity assumption is relaxed for this dimension of the model (Banker and Morey, 1986). Therefore, as the line of reasoning unfolded above is applicable to VRS technologies, estimated with convex and free-disposal hull boundaries, we need to adapt the bootstrap procedure to accommodate for the categorical variable. This can simply be done by performing the algorithm for the specific VRS frontier against which a DMU is gauged. This frontier is determined by all DMU with an equal or lower level $l \in \{1, 2, \ldots, L\}$ of the categorical variable, i.e. all DMU working in similar or harsher conditions.

[9]The efficient facet determining the value of $\hat{\theta}_{VRS}(\mathbf{x}, \mathbf{y})$ appears too often and with a fixed probability in the pseudo-samples.

The remaining question now, is which subsample size m we need to choose for each level of the categorical variable. The value of m is determined through $m = n^{\kappa}$, with $0 < \kappa < 1$. Following the approach of Simar and Zelenyuk (2007), who developed a group-wise subsampling algorithm for testing efficiency differences between L subgroups of a set of DMU, we will subsample with a value of κ being equal for all levels l of the categorical variable. That is, $m_l = n_l^{\kappa}$ for all l, with n_l = the number of observations of level $\leqslant l$.

Thus, after calculating the efficiency estimates $\hat{\theta}^l_{VRS}(\mathbf{x}, \mathbf{y})$ with equation (2.1), the following subsample bootstrap algorithm can be applied to obtain the bias-corrected estimates:

1. Generate a bootstrap sample $S^*_{m_l,b}$ for the level $l \in \{1, 2, \ldots, L\}$ by randomly drawing (independently, uniformly, and with replacement) m_l observations from the original set of n_l observations determining the empirical production possibility set for the observations of level l, i.e. all observations of level $\leqslant l$, with $m_l = \lfloor {n_l}^{\kappa} \rfloor, 0 < \kappa < 1$, and $\lfloor {n_l}^{\kappa} \rfloor$ being the largest integer smaller than ${n_l}^{\kappa}$.

2. For each observation in the level $l \in \{1, 2, \ldots, L\}$, compute the bootstrap estimate $\hat{\theta}^{*,l}_{VRS,m_l,b}(\mathbf{x}, \mathbf{y})$ using the bootstrap pseudo-sample $S^*_{m_l,b}$ from the previous step, and applying equation (2.3), with $m = m_l$.

3. For each level $l \in \{1, 2, \ldots, L\}$, repeat the above steps (1) and (2) B times and obtain bootstrap estimates for each $b = 1, \ldots, B$.

4. The resulting B bootstrap values can then be used to estimate the bias for each observation:

$$\widehat{BIAS}_B\Big(\hat{\theta}^l_{VRS}(\mathbf{x}, \mathbf{y})\Big) = \Big(\frac{m_l}{n_l}\Big)^{\frac{2}{N+M+1}} \Big[\frac{1}{B}\sum_{b=1}^{B} \hat{\theta}^{*,l}_{VRS,m_l,b}(\mathbf{x}, \mathbf{y}) - \hat{\theta}^l_{VRS}(\mathbf{x}, \mathbf{y})\Big], \qquad (2.4)$$

   with the factor $\big(\frac{m_l}{n_l}\big)^{\frac{2}{N+M+1}}$ correcting for the effect of different sample size in the original data and the bootstrap subsamples Simar and Wilson (2008).

5. The bias-corrected estimates can then be obtained by:

$$\hat{\hat{\theta}}^l_{VRS}(\mathbf{x}, \mathbf{y}) = \hat{\theta}^l_{VRS}(\mathbf{x}, \mathbf{y}) - \widehat{BIAS}_B\Big(\hat{\theta}^l_{VRS}(\mathbf{x}, \mathbf{y})\Big). \qquad (2.5)$$

We programmed the bootstrapping algorithm elaborated above in the R environment. Only for the basic DEA calculation, i.e. equation 2.2, the FEAR 2.0 package (Wilson, 2008)

was applied. Bootstrap calculations were performed with $B = 2000$ iterations. After assessing the stability in the bootstrap results with several values of κ, both for the working week and weekend datasets, we chose a value for κ equal to 0.75 (Daraio and Simar, 2007).

2.3.4 Second stage regressions

In order to gain insight in the determinants of railway traffic control efficiency, we explore a series of possible causes in the second stage of our framework. To this end, several researchers have applied second-stage regressions on efficiency scores estimated by means of DEA models with categorical variables (e.g. Teresa Balaguer-Coll and Prior, 2009; Harrison and Rouse, 2014).

We mainly follow the approach adopted by Button and Neiva (2013, 2014) in their analysis of Europe's Air Navigation Service Providers, and perform an OLS regression on the bias-corrected efficiency estimates. As the use of second stage regression methods is currently the subject of an academic debate, in which McDonald (2009), Banker and Natarajan (2008),and Simar and Wilson (2007, 2011) play a leading role, we complement this approach with a truncated regression on the bias-corrected scores, applying the single bootstrap procedure developed by Simar and Wilson (2007). Finally, in line with the recommendation of McDonald (2009) to calculate White's heteroskedastic-robust standard errors for the OLS regressions, we apply Arellano clustered standard errors for panel data (robust to heteroskedasticity and temporal serial correlation).

2.4 Data

We will demonstrate the practical applicability of our framework with a unique and rich set of intra-company data provided by Infrabel. Detailed staff rostering and operations data for relay-based signal boxes were gathered, for an 18 month period starting from January 2013 till June 2014. Given the substantial differences between the staffing levels and traffic densities in the working week and the weekend, we looked for additional insights and patterns by splitting up the monthly data in two subsets, one covering the five weekdays of the working

week (Monday to Friday), the other the weekend (Saturday and Sunday)[10]. Results and discussions reported in this paper will be based on these 2 datasets.

With the aim of implementing the efficiency analysis as an ongoing exercise, a custom Business Intelligence application code-named as the CRIPTON Business Intelligence tool was developed. The tool collected micro-data from the databases of interest, and subsequently aggregated the data to signal box level, the Decision Making Unit which is the subject of our DEA efficiency analysis. In line with the objective of closely monitoring traffic control performance, monthly datasets were generated. With the CRIPTON tool, a detailed drill-down analysis of the underlying data as well as an interactive visualization of the efficiency results were made accessible at the click of a mouse[11]. A cornerstone of this concept was the creation of a new database, linking data from the staff rostering application with data from the operational systems. The server-based tool was built in close cooperation with Infrabel's Traffic Operations department, with the specific aim of not only preparing the necessary data sets, but also verifying data quality and introducing the DEA concept in the organisation. Most importantly, the use of the Business Intelligence tool helped to unlock the full potential of the expert panel, and proved to be an important asset in the process of building the DEA-based framework and validating the empirical results.

The initial dataset generated by the CRIPTON Business Intelligence tool consisted of 101 relay-technology signal boxes. Together with the expert panel, an extensive data examination was carried out. Due to complexities inherent to the migration process, 8 signal boxes were eliminated from the sample (as they are temporarily equipped with mixed technologies, relay-based and electronic). Another 10 exhibited errors in the data, mainly in the first months of the sample, and 3 signal boxes presented local particularities which could not be modelled in the database. The list of 80 remaining signal boxes was validated by the expert panel. During the 18 months under consideration, and as a consequence of the ongoing migration towards

[10]As pointed out by an anonymous referee, this data split could be further improved by also considering public holidays as weekend days. In addition, public holidays with a large-scale shutdown of the railway system (such as 'boxing day' in the UK) could possibly lead to very poor efficiency levels, and should be analysed with care. Such cases do not occur on the Belgian network.

[11]DEA calculations were performed in R, subsequently imported in the Business Intelligence tool, and interactively visualised next to micro-level data such as the corresponding railway lines, nodes, signals, train numbers, or staff rostering details.

electronic technology signal boxes, 14 relay-technology signal boxes gradually left the sample, leading to a total of 1305 observations.

As there was no reliable data available on the shunting movements, the expert panel made an assessment of the appropriate level of shunting workload for each signal box. Thus, the sample was categorized as follows:

Table 2.1: Shunting levels (categorical variable SHUNT)

Level	**shunting workload (% of total movements)**	**# of signal boxes**	**# of monthly observations**
1	100% - 80%	24	405
2	80% - 60%	7	126
3	60% - 40%	9	141
4	40% - 20%	13	215
5	20% - 0%	27	418
Total		80	1,305

It was also decided to exclude the first level from the sample, as the shunting workload for these signal boxes was judged as being consistently close to 100% of the total movements (i.e. signal boxes in shunting yards). In doing so, the sample was further reduced to 900 observations. Table 2.2 provides the descriptive statistics of the final datasets (working week and weekends). In this table, the name of the 3 categories of environmental variables reflects the decision-making level which has authority over these variables.

2.5 Results

2.5.1 DEA results

Table 2.3 summarizes the obtained efficiency scores for the working week and weekend estimations. On average, the bias correction leads to a slight decrease in average efficiency scores of 0.017 (working week) to 0.014 (weekend). For the remainder of this paper, we will only consider the bias-corrected values.

Table 2.2: Descriptive statistics

		Working week (Mon-Fri)				Weekends (Sat-Sun)			
		Mean	St. dev.	Min	Max	Mean	St. dev.	Min	Max
1. Production process									
Input									
HOURS	*(hours worked)*	1,009	513	304	3,574	351	170	29	1,090
Output									
TRAIN	*(train movements)*	11,727	10,640	1,036	59,991	2.402	2.276	40	13,916
SHUNT	*(shunting level)*	4.028	1.087	2	5	4.028	1.087	2	5
LINES	*(line.km controlled)*	451.8	383.2	60.9	2,127.2	177.9	155.6	6.18	924.87
NODES	*(nodes controlled)*	90.9	67.9	13.6	299.0	34.9	26.7	2.4	130.0
2. Environmental variables influencing efficiency									
External decision-making (railway traffic characteristics)									
VAR	*(variability)*	1.794	0.235	1.260	2.506	1.745	0.583	1.000	7.273
DENS_SPAT	*(spatial density)*	13.831	7.928	2.386	42.252	7.451	4.576	0.250	22.969
DENS_TEMP	*(temporal density)*	0.941	0.899	0.000	4.120	0.314	0.403	0.000	2.839
TT_CHAR	*(timetable charact.)*	0.005	0.007	0.000	0.042	0.005	0.010	0.000	0.084
Internal decision-making (asset management policy)									
COMP_NET	*(network complexity)*	1.036	0.387	0.250	2.000	1.036	0.387	0.250	2.000
COMP_TRACK	*(track complexity)*	11.241	5.574	3.625	25.000	11.241	5.573	3.625	25.000
P_STATIONS	*(proportion stations)*	86.80	20.46	25.00	100.00	86.80	20.46	25.00	100.00
WORK_DENS	*(infrastr. works)*	0.223	0.463	0.000	3.222	0.153	0.331	0.000	2.675
Internal decision-making (traffic control policy)									
P_CLOSED	*(closing times)*	5.929	11.136	0.000	42.029	8.605	17.838	0.000	86.574
N_PERSONS	*(team size)*	13.657	6.383	3	42	12.387	6.004	2	40
KM_PERSON	*(centralisation)*	2.818	1.715	0.152	8.633	3.074	1.798	0.152	9.270
AVG_AGE	*(age of the staff)*	49.77	3.46	36.67	58.69	49.57	3.92	36.47	58.69
ERRORS	*(errors delays)*	0.136	0.861	0.000	25.038	0.137	0.768	0.000	18.416

Table 2.3: Efficiency scores

	Working week (Mon-Fri)				Weekends (Sat-Sun)			
	Mean	Median	St. dev.	Min	Mean	Median	St. dev.	Min
efficiency	0.664	0.682	0.215	0.260	0.574	0.483	0.237	0.159
bias	0.017	0.014	0.016	0.000	0.014	0.009	0.014	0.000
bias-corrected efficiency	0.647	0.670	0.210	0.256	0.560	0.473	0.232	0.156

Average efficiency levels may seem rather low, but this is a consequence of the unavoidable 'available time' in signal boxes, since the workload associated with the traffic volumes and the supervised infrastructures cannot always sufficiently fill each (e.g. 8-hour) working shift. In addition, as we shall see in the regression results, there are several factors not under the control of local management which significantly influence efficiency, and therefore can impede efforts to maximise efficiency. Very low efficiency scores can be observed at the periphery of the network, where few trains run on relatively short stretches of track. It should also be emphasized that, although the calculated technical efficiency scores suggest a sometimes large potential for performance improvement, major productivity and efficiency gains are only achievable through the implementation of a different technology (i.e. the migration towards electronic signal boxes). The DEA calculations do nevertheless allow senior management to look for smaller and incremental efficiency improvements, by analysing the best and worst practices across their (sometimes extensive) network, and keeping a finger on the pulse through a continuous monitoring of traffic control performance.

The results also show a strong difference between the average efficiency levels in the working week versus the weekend: mean efficiency scores drop substantially from 0.647 for the working week to 0.560 for weekend efficiency (difference of 0.087), while the median shifts from 0.670 to 0.473, i.e. minus 0.197. Although this *weekend effect* was not entirely unexpected by the Infrabel expert panel, it was now quantified for the first time, and identified as being statistically significant (Wilcoxon signed-rank test statistic: $V = 318414$, p-value $< 2.2e\text{-}16$). Correlation between working week and weekend efficiencies is positive and significant, but not extremely large: the Pearson correlation coefficient is 0.762 (p-value $< 2.2e\text{-}16$), while the Spearman coefficient equals 0.745 (p-value $< 2.2e\text{-}16$).

Taking a closer look at the gap between working week and weekend efficiency (see the histogram in figure 2.1 with the calculated efficiency difference for each signal box), we can observe that working week efficiency is not consistently larger than weekend efficiency, and that a higher weekend efficiency occurs for a substantial number of signal boxes. An in-depth analysis of some of the latter cases unravelled a series of explanations, such as modified signal box closing times, or a closer alignment of staffing levels to the traffic volumes. Traffic densities were consistently lower during the weekends. As mentioned, the factors influencing

efficiency will be examined more closely in the second stage regressions of the benchmarking framework.

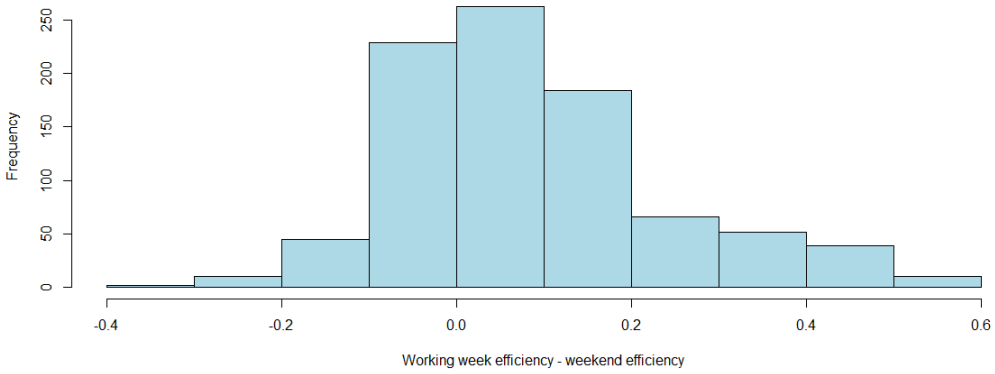


Figure 2.1: Histogram of observed efficiency differences between working week and weekend

The detailed reporting of the DEA results - which will not be disclosed here for confidentiality reasons - displays an average efficiency trend which seems to be slightly positive for working week, and stable for the weekends. Seasonal effects appear to be most clearly present during the weekends, with higher average efficiencies during the summer months (which the expert panel interpreted as a consequence of increased closing times). The observed average efficiency evolutions are a consequence of both the migration strategies (privileging the elimination of signal boxes perceived as less efficient), and tendencies related to the remaining relay-technology signal boxes (with the traffic of the eliminated signal boxes being taken over by the new electronic signal boxes). Interestingly, although the efficiency of the individual signal boxes seems to be relatively robust over time, several signals boxes display efficiency changes which, after conducting a more thorough analysis by the experts, revealed a very diverse pattern of underlying causes (e.g. slowly evolving towards best-practice through local optimisation of staff rostering plans, or gradual decrease in efficiency due to long-term changes in traffic volume).

This is an illustration of performance trends which can 'develop slowly and sometimes unevenly across different units', as indicated by Brockett et al. (1999). In order to detect and monitor these and other trends, the combination of the DEA methodology (providing a single measure of efficiency of the complex production process in the signal boxes) with the ease of

use of a Business Intelligence tool (allowing for tailored management reporting as well as an in-depth analysis by experts), can be of considerable value to decision-makers.

2.5.2 Regression results

In the second stage of the benchmarking framework, the bias-corrected efficiency scores (independent variable) are regressed against the environmental variables presented in the methodology section. A positive sign of the parameter estimates implies a positive impact of the environmental variable on technical efficiency. The results of these second-stage regressions[12] are presented in table 4. Both for the OLS and the truncated bootstrap regression, two model specifications are tested: a first model with only the traffic and asset management variables (TR_AM), and a second model TR_AM_TC including the full array of explanatory variables. Moreover, the juxtaposition of working week and weekend results allows for additional insights and robustness checks.

Bivariate correlations between independent variables did not indicate multi-collinearity problems. All variance inflations factors for the estimators of the OLS models are well below the threshold of 5, with a maximum value of 2.18. In particular, variables which might seem related at first sight (such as density and complexity, or team size and geographical centralisation) exhibit a low correlation[13]. According to the Infrabel experts, the low correlation between infrastructure complexity and traffic density can be explained by the design of the train routing across the track configuration (even at a lower traffic density, the train routing can require a more complex infrastructure, and vice versa). Also, although the concepts of team size and geographical concentration seem closely related, they are complementary dimensions (which show little correlation in our sample of relay-technology signal boxes) and should not be equated with each other. For example, the larger team sizes can also be the

[12] All calculations were carried out in R. OLS regressions were performed with the plm package. The truncated single bootstrap regressions are developed with the functions available in the FEAR 2.0 package. We also performed various robustness checks with related model specifications, as well as OLS regressions on the original efficiency estimates, all showing similar results (not reported here). As suggested by an anonymous referee, we also calculated the DEA model without weighing the train movements (see the description of the TRAIN variable). This provided similar regression results.

[13] Correlations between N_PERSONS and KM_PERSON: 0.14 in working week, 0.21 in weekends; correlations between COMP_GRID and DENS_SPAT (DENS_TEMP): 0.38 (0.22) in working week, 0.27 (0.10) in weekends.

consequence of dense traffic areas or important shunting activities, in signal boxes covering only short stretches of track.

We control for trends and seasonal effects through 2 semester dummies (representing the last six months of 2013 and the first six months of 2014). As our base methodology is the OLS regression with cluster-robust standard errors, we will mainly discuss results based on this approach. We will also focus on the most general regression model (TR_AM_TC), and highlight the most important differences and similarities between working week and weekends.

Overall, our regression results exhibit a moderately strong goodness-of-fit (adjusted R-squared: 0.63 for the working week, 0.71 for the weekend). The model TR_AM shows a more modest but still satisfactory explanatory power (adjusted R-squared: 0.42 for the working week, 0.47 for weekends). In terms of confidence intervals, the OLS model with cluster-robust standard errors generally yields more cautious results then the (Simar and Wilson, 2007) single bootstrap procedure, which does not correct for possible heteroskedasticity and serial correlation in the panel data.

First, we discuss the environmental variables representing traffic and timetable characteristics (variables not under the control of the infrastructure manager). The impact of traffic variability VAR exhibits a positive sign but is only significant during weekends (attaining significance in 3 out of the 4 regressions). An intuitive explanation by the expert panel was that weekends typically display a larger difference between day-time and night-time traffic volumes. Although more research is needed on this aspect (e.g. through more precise modelling of traffic variability), it would appear that signal boxes are able to cope with traffic decline during the weekends, e.g. through reduced night shifts. Even though statistical significance is not achieved in the working week models, the positive sign of the regression coefficient could also point at the adaptability of the signal boxes to follow traffic variations.

The variables exploring the influence of traffic density demonstrate the anticipated effect on efficiency levels, although consistent statistical significance is only attained in the working week. The higher traffic densities during the working week could explain this dissimilarity with the weekends. As expected, the spatial traffic density DENS_SPAT has a positive impact on railway traffic efficiency, while the temporal traffic density DENS_TEMP exerts a negative

influence. The last variable related to traffic and timetable, train connections and changes in rolling stock and crew at station platforms TT_CHAR, is not significant except for the two truncated regressions in the weekends (and carries an unexpected positive sign). A possible explanation for this counterintuitive result is that the variable is a poor proxy for the characteristics of the timetable it is intended to operationalise (e.g. number of delays generated by train connections, instead of the true number of connections).

Turning next to the group of variables related to railway infrastructure asset management, the network complexity COMP_NET is not consistently significant across the regression models. However, highly significant negative effects of track layout complexity (COMP_TRACK) can be observed. As the COMP_TRACK variable was proxied by the number of signals per node, we need to interpret this result with some caution. The variable does account for the number of signals, but may not necessarily fully reflect additional complexity parameters such as the number of switches or the possible routes in the track configuration. We refer to Landex and Jensen (2013) for a series of track complexity measures which, however, require much more detailed data, such as the number of conflicting train routes. In line with the expectations of the expert panel, the final complexity variable P_STATIONS, i.e. the proportion of stations in the network, exerts a positive influence (and is clearly significant). The negative impact of the density of infrastructure works WORK_DENS is only significant in the weekend models (3 out of 4 regressions). As infrastructure works mainly take place during the night or in the weekends, this result was not entirely unanticipated.

The final group of environmental variables consists of parameters under the control of the central management responsible for the signal boxes, but which are beyond the discretionary power of local management. The percentage of signal box closing times (variable P_CLOSED), is confirmed as highly significant throughout all models, with a positive impact both for the working week and the weekend. The factor team size (variable N_PERSONS) also positively influences efficiency, and is highly significant in all models. Team size is closely linked to the input variable HOURS. Therefore this result must be interpreted as the impact of scale on efficiency, after allowing for scale effects when determining the production frontier (as we applied the DEA Variable Returns to Scale model). In other words, our results indicate that a larger scale – in terms of team size – allows signal boxes to move closer to the production

frontier, and hence increase the efficiency of their operations.

Regression results also identify a clear positive and significant impact of the variable KM_PERSON, which reflects the degree of geographical centralisation. As this result is in accordance with previous rail and air traffic control research (e.g. International Union of Railways, 2002; Civity management consultants, 2013; Button and Neiva, 2013, 2014), and is also confirmed by the current migration strategies towards centralised traffic control centres, this provides us with further confidence in our findings. In addition, even though the 95 % confidence intervals of the working week and weekend slightly overlap for the OLS estimations, the higher regression coefficient for the weekend results could also point - ceteris paribus - at a higher leverage of geographical centralisation on the weekend efficiencies. This could be a consequence of the lower traffic volumes, which allows for a higher coverage of railway line capacity per person. Another explanation, applicable in some cases, is the partial closing of signal boxes in the weekends (which remains uncaptured by the data). Evidently, more research is needed to investigate this particular phenomenon. The last two environmental variables, average age AVG_AGE and human errors ERRORS are all insignificant[14], as well as the semester dummies, and are therefore not reported. Including monthly dummies or a time trend provided similar results.

[14]Results are robust to omission of these variables.

Table 2.4: Regression results on bias-corrected efficiency estimates (working week)

	Working week (Mon-Fri)			
	OLS (robust SE)[a]	OLS (robust SE)[a]	trunc. bootstrap[b]	trunc. bootstrap[b]
	TR_AM	TR_AM_TC	TR_AM	TR_AM_TC
Constant	1.142***	0.435*	1.377***	0.414***
	(0.840, 1.444)	(-0.012, 0.881)	(1.222,1.507)	(0.247,0.582)
External decision-making (railway traffic characteristics)				
VAR	0.025	0.068	0.020	0.081***
(variability)	(-0.130, 0.180)	(-0.046, 0.181)	(-0.042,0.083)	(0.035,0.127)
DENS_SPAT	0.006**	0.005**	0.007***	0.004***
(spatial density)	(0.001, 0.011)	(0.001, 0.010)	(0.005,0.009)	(0.002,0.006)
DENS_TEMP	-0.059***	-0.034*	-0.072***	-0.025***
(temporal density)	(-0.101, -0.016)	(-0.072, 0.004)	(-0.087,-0.052)	(-0.038,-0.011)
TT_CHAR	0.015	-0.568	-0.919	-0.933
(timetable characteristics)	(-4.397, 4.426)	(-4.091, 2.955)	(-2.902,1.036)	(-2.327,0.501)
Internal decision-making (asset management policy)				
COMP_NET	-0.008	-0.054	-0.013	-0.057*
(network complexity)	(-0.136, 0.120)	(-0.146, 0.038)	(-0.051,0.027)	(-0.090,-0.024)
COMP_TRACK	-0.015***	-0.011***	-0.016***	-0.011***
(track complexity)	(-0.022, -0.008)	(-0.017, -0.005)	(-0.018,-0.013)	(-0.013,-0.009)
P_STATIONS	-0.005***	-0.004***	-0.007***	-0.005***
(proportion of stations)	(-0.006, -0.003)	(-0.005, -0.002)	(-0.008,-0.006)	(-0.006,-0.004)
WORK_DENS	0.018	0.012	0.018	0.009
(infrastructure works)	(-0.027, 0.063)	(-0.026, 0.050)	(-0.011,0.050)	(-0.011,0.032)
Internal decision-making (traffic control policy)				
P_CLOSED		0.009***		0.012***
(closing times)		(0.005, 0.014)		(0.011,0.014)
N_PERSONS		0.009***		0.011***
(team size)		(0.006, 0.013)		(0.009,0.013)
KM_PERSON		0.034***		0.045***
(geographical centralisation)		(0.014, 0.054)		(0.038,0.051)
R2 (adjusted R2)	0.424 (0.418)	0.645 (0.633)		

[a] heteroskedastic and temporal serial correlation robust standard errors, Arellano (1987)

[b] Simar and Wilson (2007) single truncated bootstrap

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$; 95 % CI between brackets.

Table 2.5: Regression results on bias-corrected efficiency estimates (weekends)

	Weekends (Sat-Sun)			
	OLS (robust SE)[a]	OLS (robust SE)[a]	trunc. bootstrap[b]	trunc. bootstrap[b]
	TR_AM	TR_AM_TC	TR_AM	TR_AM_TC
Constant	1.181***	0.630**	1.325***	0.617***
	(0.911, 1.451)	(0.143, 1.118)	(1.216,1.393)	(0.491,0.737)
External decision-making (railway traffic characteristics)				
VAR	0.039	0.052**	0.040***	0.056***
(variability)	(-0.043, 0.122)	(0.005, 0.099)	(0.015,0.062)	(0.038,0.072)
DENS_SPAT	0.007	0.006	0.007***	0.006***
(spatial density)	(-0.003, 0.017)	(-0.002, 0.014)	(0.003,0.010)	(0.003,0.008)
DENS_TEMP	-0.021	-0.035	-0.019	-0.033***
(temporal density)	(-0.112, 0.070)	(-0.107, 0.036)	(-0.054,0.019)	(-0.056,-0.006)
TT_CHAR	2.835	1.133	2.604***	1.020**
(timetable characteristics)	(-1.254, 6.924)	(-1.272, 3.538)	(1.169,3.852)	(0.098,1.890)
Internal decision-making (asset management policy)				
COMP_NET	-0.049	-0.028	-0.047**	-0.026
(network complexity)	(-0.163, 0.065)	(-0.104, 0.048)	(-0.080,-0.008)	(-0.056,0.004)
COMP_TRACK	-0.017***	-0.010***	-0.017***	-0.009***
(track complexity)	(-0.024, -0.009)	(-0.015, -0.004)	(-0.019,-0.014)	(-0.011,-0.008)
P_STATIONS	-0.006***	-0.005***	-0.007***	-0.005***
(proportion of stations)	(-0.008, -0.003)	(-0.006, -0.003)	(-0.008,-0.006)	(-0.005,-0.004)
WORK_DENS	-0.076**	-0.029	-0.084***	-0.036**
(infrastructure works)	(-0.134, -0.018)	(-0.079, 0.021)	(-0.118,-0.042)	(-0.060,-0.009)
Internal decision-making (traffic control policy)				
P_CLOSED		0.004***		0.005***
(closing times)		(0.002, 0.006)		(0.004,0.006)
N_PERSONS		0.009***		0.009***
(team size)		(0.004, 0.015)		(0.008,0.011)
KM_PERSON		0.062***		0.069***
(geographical centralisation)		(0.047, 0.077)		(0.063,0.075)
R2 (adjusted R2)	0.478 (0.472)	0.725 (0.712)		

[a] heteroskedastic and temporal serial correlation robust standard errors, Arellano (1987)

[b] Simar and Wilson (2007) single truncated bootstrap

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$; 95 % CI between brackets.

2.6 Conclusions

In this paper, we presented a first of many steps in the new and in our opinion promising research field of railway traffic control efficiency. Drawing on related research as well as railway expert knowledge, we constructed a DEA-based benchmarking framework which assesses and explains the relative efficiency of traffic control in signal boxes. In a first stage, the framework estimates the technical efficiency of the production process, and keeps close track of average and individual performance trends over time. The efficiency scores are bias-corrected with a DEA subsample bootstrap algorithm, which we adapted to accommodate for DEA models with a categorical variable. The impact of several determinants of efficiency is examined in second-stage regressions. We demonstrated the practical applicability of the developed framework on a unique and rich 18-month dataset of Infrabel's' relay-technology signal boxes. Aiming to uncover additional insights, our calculations were performed on two subsets containing working week and weekend data. The analysis was supported by the development and implementation of a custom Business Intelligence application. This tool proved to be an important asset, not only as a managerial instrument, but also during the process of building and validating the DEA framework.

As the basic principles of railway operation are similar across Europe (Pachl, 2009, preface), and as the DEA methodology relies on a minimum of a priori assumptions, we are confident that our framework can be adopted by other infrastructure managers. It can be applied as a decision-support tool for senior management, internally benchmarking the entire network or specific sub-regions. The single overall measure of efficiency obtained through the DEA calculations can act as a guide to pinpoint the best, good and worst practices throughout the examined area. Especially for large networks with an extensive number of signal boxes (such as the French or British, see the introduction) this can deliver powerful management insight. If the goal is to consistently inform the decision makers on efficiency trends, the tool should preferably be implemented as an ongoing exercise, supported by advanced reporting and analysis software. As the development of such a performance measurement system can consume important time and resources, it should be approached as a long-term and sustainable project, with considerable academic input (or sufficient internal capabilities) and an appropriate project management structure. And finally, but most importantly, it

needs continuous support from the management involved.

Two sets of policy recommendations for infrastructure managers can be drawn from our empirical results. First, oriented towards the asset management component of railway infrastructure, our second-stage results suggest a significant influence of track layout complexity on efficiency. This could imply that an asset management strategy, aiming for 'lean infrastructure' (International Union of Railways, 2002) is not only reducing asset maintenance cost, but also has positive effects on traffic control efficiency. At Infrabel, the reduction of infrastructure complexity (while still maintaining the same levels of capacity and flexibility in handling the traffic volumes) is a long-term and ongoing process, integrated in the infrastructure renewal program.

A second set of conclusions is relevant for railway traffic control policy. Our DEA efficiency results show that average efficiency levels clearly and significantly drop during the weekend, thus confirming the intuition that the lower weekend traffic volumes decrease efficiency. More surprisingly however, at an individual level, a higher weekend efficiency can be observed for a substantial number of signal boxes (even though traffic densities consistently remained lower during the weekends). These diverging 'weekend effects' can further assist senior management in identifying and analysing their best and good practices, which may be different in weekends compared to the working week.

Based on the second stage regression results, further policy recommendations regarding railway traffic control can be put forward. First, geographical centralisation and a higher team size clearly and significantly improve efficiency levels. Although both concepts seem closely related, they are complementary dimensions (which show little correlation in our sample of relay-technology signal boxes) and should not be equated with each other. For example, the larger team sizes can also be the consequence of dense traffic areas or important shunting activities, in signal boxes covering only short stretches of track. As indicated in the international benchmarking report from the UK Office of Rail Regulation (Civity management consultants, 2013), larger team sizes allow for a more flexible and closer alignment of the working shifts to the hourly traffic profile, and as such offer the potential to increase efficiency. Infrastructure managers should therefore complement the beneficial effects of geographical centralisation with the optimisation of their staff rostering, an exercise which can be leveraged by larger

team sizes. The current migration strategies across Europe, aiming for fewer and larger signal boxes, provide this opportunity to further improve on efficiency through optimised resource planning (see ibid.).

Second, the opening and closing of infrastructure for operation provides a significant lever for increasing efficiency. Although the power to change opening times can be restricted by operational constraints (such as train paths demanded by railway undertakings), it is a key parameter to improve efficiency. It does not require extensive investment budgets, and has the potential to deliver results in a relatively short time span. A practical implementation of this measure could be supported by a thorough and systematic monitoring of areas with very weak traffic volumes at the early or late hours of the day. Slowly changing traffic volumes (e.g. in freight traffic) can then act as a trigger to examine the opening hours of the signal boxes along the affected railway axes, or consider a partial closing of the infrastructure. In addition, as put forward in International Union of Railways (2002), shunting operations could be analysed and bundled into fewer hours.

Although the practical application of our framework was demonstrated on relay-technology signal boxes, we expect the results to be generalizable to other signal box technologies, and to railway networks or regions with a comparable range of traffic density and infrastructure complexity. An element however not considered in our study, is the automation of the signalling activities through Automatic Route Setting[15] (ARS). The automation can provide an additional lever for efficiency improvement. At Infrabel, ARS is currently being rolled out in the electronic signal boxes, and is introduced with the objective of not only further enhancing the efficiency but also the quality of traffic control. We refer the interested reader to Hayden-Smith (2013) for a more detailed discussion on the impact of ARS on signaller workload. In this interview-based analysis, areas with high traffic density and a higher infrastructure complexity are expected to still require considerable manual intervention (a consequence of knock-on delays passed from one train to another).

In order to further improve the DEA model, our next research efforts will directed towards

[15] Automatic Route Setting (ARS): the automatic setting of a train route when a train approaches a signal (Pachl, 2009, p. 228). ARS software is developed for electronic signal boxes, and is therefore not considered in our analysis of relay-technology signal boxes.

the internal process flows in the signal boxes. In conventional DEA, the production unit under consideration (e.g. the signal box) is modelled as a 'black box' which transforms the inputs into outputs. By including information on the internal production process, the efficiency results can provide additional managerial insights. One approach currently under consideration is the novel DEA-based methodology developed by Cherchye et al. (2013), which incorporates expert knowledge on the process flows into the benchmarking models.

Acknowledgements

The authors would like to thank Infrabel for funding this research, and providing the necessary data and expertise. Special thanks to Jean-Claude Dechef, Didier Dehandschutter, Freddy Lhermitte, Luc D'Hollander, and Sabine Verboven. It is however to be noted that the views expressed in this paper are those of the authors and do not necessarily reflect the opinions of Infrabel. We are also most grateful to Marijn Verschelde for the methodological feedback. The international perspective offered by Michael Robson is also greatly appreciated. Finally, we would like to thank the participants of the DEA 2013 Workshop on Productivity, Regulation & Transportation in Jerusalem (Hebrew University) for the valuable first discussions on the approach, and three anonymous referees for their constructive and insightful comments. Any remaining errors are the sole responsibility of the authors.

References

Andersson, M. (2008). Marginal railway infrastructure costs in a dynamic context. *European Journal of Transport and Infrastructure Research*, 4(8).

Banker, R. D., Charnes, A., and Cooper, W. W. (1984). Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management science*, 30(9):1078–1092.

Banker, R. D. and Morey, R. C. (1986). The use of categorical variables in data envelopment analysis. *Management science*, 32(12):1613–1627.

Banker, R. D. and Natarajan, R. (2008). Evaluating contextual variables affecting productivity using data envelopment analysis. *Operations research*, 56(1):48–58.

Boussofiane, A., Dyson, R. G., and Thanassoulis, E. (1991). Applied data envelopment analysis. *European Journal of Operational Research*, 52(1):1–15.

Brockett, P. L., Golany, B., and Li, S. (1999). Analysis of intertemporal efficiency trends using rank statistics with an application evaluating the macro economic performance of oecd nations. *Journal of Productivity Analysis*, 11(2):169–182.

Button, K. and Neiva, R. (2013). Single european sky and the functional airspace blocks: Will they improve economic efficiency? *Journal of Air Transport Management*, 33:73–80.

Button, K. and Neiva, R. (2014). Economic efficiency of european air traffic control systems. *Journal of Transport Economics and Policy*, 48(1):65–80.

Cherchye, L., Rock, B. D., Dierynck, B., Roodhooft, F., and Sabbe, J. (2013). Opening the "black box" of efficiency measurement: input allocation in multioutput settings. *Operations Research*, 61(5):1148–1165.

Civity management consultants (2013). International benchmarking of network rail's operations and support functions expenditure. *Department for Transport and Office of Rail Regulation; London.*

Coelli, T. J., Rao, D. S. P., O'Donnell, C. J., and Battese, G. E. (2005). *An introduction to efficiency and productivity analysis*. Springer Science & Business Media.

Cooper, W. W., Seiford, L. M., and Zhu, J. (2011). Data envelopment analysis: History, models, and interpretations. In *Handbook on Data Envelopment Analysis*, pages 1–39. Springer.

Cowie, J. and Loynes, S. (2012). An assessment of cost management regimes in british rail infrastructure provision. *Transportation*, 39(6):1281–1299.

Daraio, C. and Simar, L. (2007). *Advanced robust and nonparametric methods in efficiency analysis: Methodology and applications*. Springer Science & Business Media.

EUROCONTROL Performance Review Commission (2011). Econometric cost-efficiency benchmarking of air navigation service providers. *EUROCONTROL, Brussels*.

European Directive 2012/34/EU (2012). Directive 2012/34/EU of the European Parliament and of the Council of 21 november 2012 establishing a single european railway area. *Official Journal of the European Union, 55*.

Golany, B. and Roll, Y. (1989). An application procedure for DEA. *Omega*, 17(3):237–250.

Growitsch, C. and Wetzel, H. (2009). Testing for economies of scope in european railways: an efficiency analysis. *Journal of Transport Economics and Policy*, 43(1):1–24.

Hansen, I. A., Wiggenraad, P., and Wolff, J. (2013). Performance analysis of railway infrastructure and operations. In *WCTR 2013: 13th World Conference on Transport Research, Rio de Janeiro, Brazil, 15-18 July 2013*.

Harrison, J. and Rouse, P. (2014). Competition and public high school performance. *Socio-Economic Planning Sciences*, 48(1):10–19.

Hayden-Smith, N. (2013). The future of signaller workload assessments in automated world. *Dadashi, N., Scott, A., Wilson, JR, Mills, A.(eds.)*, pages 419–426.

Holder, S., Veronese, B., Metcalfe, P., Mini, F., Carter, S., and Basalisco, B. (2006). Cost benchmarking of air navigation service providers: A stochastic frontier analysis. *Nera Economic Consulting, London*.

International Union of Railways (2002). Infracost-the cost of railway infrastructure. *Final Report, Paris*.

Johansson, P. and Nilsson, J.-E. (2004). An economic analysis of track maintenance costs. *Transport Policy*, 11(3):277–286.

Kennedy, J. and Smith, A. S. (2004). Assessing the efficient cost of sustaining britain's rail network: Perspectives based on zonal comparisons. *Journal of Transport Economics and Policy*, 38(2):157–190.

Kneip, A., Simar, L., and Wilson, P. W. (2008). Asymptotics and consistent bootstraps for DEA estimators in nonparametric frontier models. *Econometric Theory*, 24(06):1663–1697.

Landex, A. and Jensen, L. W. (2013). Measures for track complexity and robustness of operation at stations. *Journal of Rail Transport Planning & Management*, 3(1):22–35.

Löber, G. and Staat, M. (2010). Integrating categorical variables in data envelopment analysis models: A simple solution technique. *European Journal of Operational Research*, 202(3):810–818.

McDonald, J. (2009). Using least squares and tobit in second stage DEA efficiency analyses. *European Journal of Operational Research*, 197(2):792–798.

McNulty, S. R. (2011). Realising the potential of gb rail: final independent report of the rail value for money study. *Department for Transport and Office of Rail Regulation, London.*

Merkert, R. and Nash, C. A. (2013). Investigating european railway managers' perception of transaction costs at the train operation/infrastructure interface. *Transportation Research Part A: Policy and Practice*, 54:14–25.

Merkert, R., Smith, A. S., and Nash, C. A. (2010). Benchmarking of train operating firms–a transaction cost efficiency analysis. *Transportation Planning and Technology*, 33(1):35–53.

Mouchart, M. and Simar, L. (2002). Efficiency analysis of air controllers: first insights. *Consulting report*, 202.

Ozbek, M. E., de la Garza, J. M., and Triantis, K. (2009). Data envelopment analysis as a decision-making tool for transportation professionals. *Journal of Transportation Engineering*, 135(11):822–831.

Pachl, J. (2009). *Railway operation and control*. VTD Rail Publishing, Mountlake Terrace (USA).

Pellegrini, P. and Rodriguez, J. (2013). Single european sky and single european railway area: A system level analysis of air and rail transportation. *Transportation Research Part A: Policy and Practice*, 57:64–86.

Simar, L. and Wilson, P. W. (2007). Estimation and inference in two-stage, semi-parametric models of production processes. *Journal of econometrics*, 136(1):31–64.

Simar, L. and Wilson, P. W. (2008). Statistical inference in nonparametric frontier models: recent developments and perspectives. *The Measurement of Productive Efficiency (H. Fried, CAK Lovell and SS Schmidt Eds), Oxford University Press, Inc*, pages 421–521.

Simar, L. and Wilson, P. W. (2011). Two-stage DEA: caveat emptor. *Journal of Productivity Analysis*, 36(2):205.

Simar, L. and Zelenyuk, V. (2007). Statistical inference for aggregates of farrell-type efficiencies. *Journal of Applied Econometrics*, 22(7):1367–1394.

Smith, A., Wheat, P., and Smith, G. (2010). The role of international benchmarking in developing rail infrastructure efficiency estimates. *Utilities policy*, 18(2):86–93.

Smith, A. S. (2012). The application of stochastic frontier panel models in economic regulation: Experience from the european rail sector. *Transportation Research Part E: Logistics and Transportation Review*, 48(2):503–515.

Smith, A. S. and Wheat, P. (2012). Estimation of cost inefficiency in panel data models with firm specific and sub-company specific effects. *Journal of Productivity Analysis*, 37(1):27–40.

Teresa Balaguer-Coll, M. and Prior, D. (2009). Short-and long-term evaluation of efficiency and quality. an application to spanish municipalities. *Applied Economics*, 41(23):2991–3002.

Tulkens, H. and Vanden Eeckaut, P. (2006). *Nonparametric efficiency, progress and regress measures for panel data: Methodological aspects*. Springer.

Wheat, P. and Smith, A. S. (2008). Assessing the marginal infrastructure maintenance wear and tear costs for britain's railway network. *Journal of Transport Economics and Policy*, 42(2):189–224.

Wilson, P. W. (2008). Fear: A software package for frontier efficiency analysis with r. *Socio-economic planning sciences*, 42(4):247–254.

Chapter 3

SHIFT WORK, FATIGUE RISK, AND HUMAN ERROR

Abstract

Fatigue is a major contributor to transportation accidents. Shiftworkers are particularly prone to fatigue, and organizations increasingly rely on fatigue risk models to evaluate work schedules. In this paper, we empirically examine the relationship between fatigue risk and human errors in railway traffic control. Despite their safety-critical role, research on railway traffic controller fatigue has remained limited. We evaluate the predictive validity of a commonly used fatigue risk tool (the Risk Index), and investigate the effect of additional risk factors (age, gender, part-time work, and day-of-week). In close cooperation with Belgian railways, we analyze a unique full year dataset, containing more than 11,000 work shifts. By adopting a Tobit regression for censored data, we account for work shifts with zero error occurrence. Our results validate the applied fatigue risk model under real-world circumstances, and reveal risk predictors above and beyond shift schedule design: significant day-of-week effects are observed. The probability of making at least one error is highest on Saturdays (+ 6 percentage points compared to Mondays), and lowest on Tuesdays, Wednesdays and Thursdays. Thus, our results suggest that safe work schedule design should also take into account the day of the week, and not exclusively rely on fatigue risk scores.

Keywords:Railway; Traffic Control; Human Factor; Fatigue; Tobit; Day-of-week.

The work in this chapter is co-authored by Johan Christiaens.

3.1 Introduction

Human fatigue is a major contributing factor in transportation accidents. Fatigue not only leads to the risk of dozing off or falling asleep, but can also result in decreased attention, slower reaction times, or memory lapses. Åkerstedt (2000) describes fatigue as 'the largest identifiable and preventable cause of accidents in transport operations', causing an estimated 15 to 20% of all accidents. In its 'Most Wanted List', the US National Transportation Safety Board (2017) identifies fatigue as a top 10 safety issue, affecting all modes of transportation.

Shiftworkers are particularly prone to fatigue problems. Improved scheduling of shiftwork has been recognized as a main countermeasure against fatigue risk (Darwent et al., 2015; Anund et al., 2015). Nowadays, several so-called biomathematical or fatigue risk models are available to evaluate the fatigue levels of work schedules. With the advent of faster computing in the 1990s, the use of these fatigue risk models - initially rooted in aviation research - has expanded to other safety-critical environments (French and Neville, 2012). Organizations increasingly rely on fatigue risk tools to manage the impact of shift roster design on human error (Gander et al., 2011; Dawson et al., 2011; Darwent et al., 2015; Dawson et al., 2017). Fatigue risk models allow to quantitatively assess the fatigue or risk associated with a given shift schedule, and are offered in interactive software (for an overview, see Mallis et al., 2004; Civil Aviation Safety Authority, 2014. As such, they provide important and relevant information for safety management purposes.

In the railway industry, fatigue research has mainly focused on train drivers. Despite their safety-critical role, there is only a limited number of studies involving railway traffic controllers or 'train dispatchers' (Dorrian et al., 2011). Railway traffic control mainly consists of authorizing train movements through signaling, making real-time dispatching decisions to mitigate delays, and ensuring safety on the network. Railway traffic controllers work around the clock, in 'traffic control centres' or 'signal boxes' (Pachl, 2009). Although scarce in numbers, research on railway traffic controller fatigue and sleep was undertaken world-wide: in Europe (e.g. Härmä et al., 2002; Sallinen et al., 2005; Cotrim et al., 2017), the US (e.g. Popkin et al., 2001; Gertler and Viale, 2007; Raslear et al., 2013), and Australia (Dorrian et al., 2011). The paucity of the research stands in contrast with the prevalence of traffic controller fatigue in US railways, where traffic controllers (dispatchers) and train staff (train

and engine workers) were found to exhibit the highest exposure to fatigue (Gertler et al., 2013). In the UK, an investigation by the Rail Safety and Standards Board (Rail Safety and Standards Board, 2015) revealed that fatigue-related incidents were mainly attributable to train drivers, followed however by traffic controllers.

The purpose of this paper is therefore to contribute to the under-researched area of railway traffic control fatigue and safety. Using a unique and rich dataset from Belgian railways, we empirically examine the relationship between human error occurrence and the fatigue risk predictions of the (Folkard et al., 2007). This fatigue risk model is issued by the UK Health and Safety Executive (HSE) in a freely available 'fatigue and risk calculator'. Together with its counterpart in the HSE tool, the Fatigue Index (Spencer et al., 2006), it is thought to be the most widely used fatigue risk model in the UK rail industry (Rail Safety and Standards Board, 2015). Our research addresses two research questions, relevant to both science and practice. Firstly, it examines the predictive validity of the Risk Index in a railway traffic control setting. As such, it provides the first Risk Index validation study in a real-life environment, more particularly in transportation. The relevance of this research question not only lies in the potential validity issues of the current fatigue risk models (as mentioned in the literature, see e.g. Dawson et al., 2011), but also in the daily use of the Risk Index in the UK rail sector. Secondly, our research seeks to uncover additional contributing factors to human error, i.e. risk predictors above and beyond shift schedule design. More specifically, we examine the effect of the day of the week, traffic controller age and gender, and part-time work.

In close cooperation with railway experts from Infrabel, the company managing Belgium's railway infrastructure[1], we analyze the work shifts in computerized Traffic Control Centres (control rooms). We apply correlations to evaluate the predictive validity of the Risk Index, and regressions to analyze the impact of the additional risk contributing factors. Our results not only validate the applied fatigue risk model (the Risk Index) in a railway traffic control environment, but also reveal a significant 'day-of-week' effect on human error, during and surrounding the weekend.

[1]Infrabel is the government-owned corporation that runs the Belgian railway infrastructure and is one of the key players in the Belgian mobility sector. Its core activities are asset management (building, maintaining and renewing the infrastructure) and traffic control.

The remainder of this paper is structured as follows. Section 2 presents further background on the fatigue risk models. Section 3 presents the research design, data and statistical procedures for the empirical analysis. Section 4 reports and discusses the empirical results. Conclusions are set out in the final section.

3.2 Fatigue risk models: background

Fatigue risk models can be categorized in two main groups (Dawson et al., 2011). The so-called one step models directly apply sleep-wake data, such as work and sleep diaries and/or wrist actigraph registrations, as an immediate input for fatigue estimations. Two-step models first estimate an average sleep-wake pattern on the basis of a given work schedule, before engaging in the second step, i.e. predicting fatigue levels. These models have a distinct practical advantage in the real world, as staff work schedules are readily available information in workplace settings (Fletcher and Dawson, 2001; Dean et al., 2007). However, the two-step estimation procedure induces additional variance in the fatigue estimations, which can lead to a decrease in its predictive ability. Nonetheless, there has been little research on the statistical reliability of two-step models, which can cast doubts on their validity in real-world settings (Dawson et al., 2011). In addition to potential issues of predictive validity, a second parallel concern raised in the literature is the application of fatigue risk models to operational settings (Friedl et al., 2004; Dean et al., 2007; Di Milia et al., 2011; Lerman et al., 2012). Clearly, it is imperative to take into account the industry-specific nature of the work tasks and their circumstances.

In our empirical analysis, we apply the Risk Index model (Folkard and Lombardi, 2004, 2006; Folkard et al., 2006; Spencer et al., 2006; Folkard et al., 2007) to estimate the risk of human errors associated with shiftwork. The Risk Index produces an assessment of the risk associated with, a given work schedule. For each work shift, it estimates the risk of occurrence of an incident or accident during the shift. The obtained risk score is relative to a typical two-day, two-night, four days off schedule with 12-hour shifts, which has an average risk score of 1. A work shift with a Risk Index of 1.25 for example, consequently has a 25% higher risk compared to this normalized 12-hour shift schedule. The risk model mainly relies on inputs of (i) the sequence of previous work and recovery days (the cumulative component

of the index), (ii) start time and length of the shift (the duty timing component), and (iii) information on workload, attention, and breaks (the job type/breaks component). The tool is based on mathematical modeling of trends in accidents and injuries, instead of relying on intermediate variables such as fatigue or alertness. It is easily understood and has a high face validity (Folkard and Lombardi, 2004). An independent study by Greubel and Nachreiner (2013) has validated the Risk Index. Their study is however based on an internet survey on working hours and occupational accidents, and is therefore not directly transposable to the real-world traffic control environment of our analysis.

Importantly, following a suggestion by Greubel and Nachreiner (2013), we specifically focus on day of the week effects by factoring in day-of-week dummy variables in our regression model. This not only contributes to the railway traffic controller fatigue and safety literature, but also to the scant body of fatigue research on day of the week effects (Monk and Wagner, 1989; Fletcher and Dawson, 2001; Brogmus, 2007; Wirtz et al., 2011; Marucci-Wellman et al., 2016). More in general, Dawson et al. (2011) identify psychosocial determinants (such as social demands during non-work periods, e.g. in the weekends) as one of the main limitations of current fatigue risk models, and recommend to incorporate these non-biological factors in future versions. Social activities can impede sleep and recovery - especially outside of laboratory settings - and can therefore influence error probability. Our regression analysis also considers independent variables capturing age, gender, and part-time work, i.e. individual factors which are not taken into account by the Risk Index model.

3.3 Data and methodology

3.3.1 Research design and datasets

In close cooperation with railway experts from Infrabel, we analyze intra-company data from 11 computerized Traffic Control Centres in Belgium, over a 12-month period stretching from November 2015 until November 2016. Shiftwork in the Traffic Control Centres is standardized through non-overlapping 8-hour shifts, starting at 06:00, 14:00 and 22:00 (i.e., early, late, and night shift). However, local management has the authority to freely organize and adapt the work schedules to balance their needs with individual shift preferences (e.g. change direction and speed of shift rotation, distribution of rest days). Our data is extracted from

the actual roster, and as such takes into account modifications such as unscheduled absences or shift swapping.

Figure 3.1 summarizes the adopted research design, with the data sources, datasets, and applied statistical procedures. All data is collected and validated trough a custom-built Business Intelligence tool, deployed at the Infrabel headquarters[2]. The application lays a cornerstone of our empirical analysis: the link between staff rostering data and the corresponding data from the operational systems. The Business Intelligence tool generates datasets at two aggregation levels. A first dataset contains observations on each separate work shift, for each day of the 12-month period, aggregated over the 11 Traffic Control Centres. A second dataset contains the same information, but is disaggregated by each of the 11 centres. Moreover, given the differences in train timetable structure for the working week and the weekend, and following Roets and Christiaens (2015) in their evaluation of Belgian railway traffic control efficiency, we perform an additional split of the two datasets (Monday - Friday, respectively Saturday - Sunday data). Our analysis focuses on the full week disaggregated data samples, and considers the other correlation and regression results as robustness checks. The full week provides the largest sample size for our analysis. In addition, the aggregated results could suffer from aggregation bias due to the loss of information (by grouping the Traffic Control Centre data).

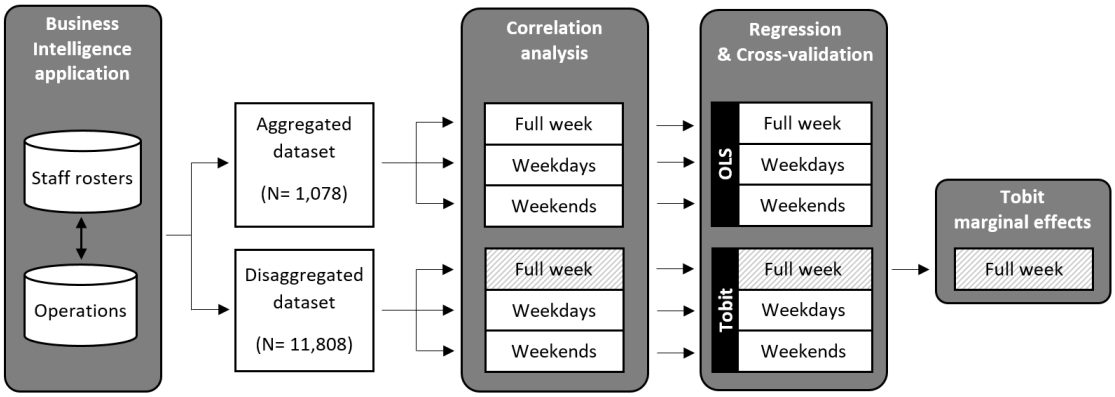


Figure 3.1: Data sources, datasets, and statistical procedures

[2]The server-based Business Intelligence software is QlikView, and is linked with the rostering and traffic control databases through a direct connection.

We apply correlation analysis to evaluate the predictive validity of the Risk Index, and regression models to analyze the impact of additional risk contributing factors. All statistical calculations are performed using R software version 3.3.2. Table 3.1 describes the variables applied in the empirical analysis. To build the datasets, we first count the human errors detected by the computerized traffic control system. The errors consist of relatively frequent but non safety-critical task errors, which can be categorized as attention failures Reason (1990). Examples are lateness in changing the train routing, misordering track signal commands, and mistyping train or track numbers. Each time an error occurs, the computer system warns the user by means of an on-screen message. The error messages are archived by the traffic control system, and can be retrieved and counted for analytic purposes. At present, the detected errors cannot be directly linked to the individual traffic controller. Therefore, we calculate the total error frequency for the entire traffic control team. After discussion and in agreement with the Infrabel railway experts, we control for exposure by dividing the registered errors by the traffic volume, and as such obtain a transparent error rate. Traffic volume is defined by the number of track signals passed by each train movement, during the work shift in question, and is divided by 1,000. Next, the Folkard et al. (2007) Risk Index is calculated for each individual traffic controller and each work shift. To analyze the correlation with the error rates, we calculate the average Risk Index for each work shift (aggregated datasets) and for each Traffic Control Centre (disaggregated datasets).

Table 3.1: Variable description

Type	Description	Correlation	Regression
Error rate	Number of human errors per 1000 train movements	X	X
Shift work risk	Average Risk Index	X	X
Day-of-week	Dummy variable for day of the week		X
Individual	Average age		X
Individual	% male traffic controllers		X
Individual	% part-time workers		X
Operational	% automatically signaled movements		X
Operational	Dummy variable for Traffic Control Centre		X
Month	Dummy variable for month of the year		X

For the purpose of the regression analysis, the datasets are further enriched with several variables from the rostering and traffic control databases. First, aiming for a statistical analysis of day-of-week effects, we add 6 dummy variables for the day of the week, with Monday serving as the reference value in the full week and weekday datasets, and Sunday in the weekend datasets. The day of the week represents the starting day for the 8-hour shift (and thus covers a period from 06:00 hours on that day, until 06:00 hours on the next day). Second, reflecting the individual characteristics of the traffic controllers, we add additional information from the rostering database. We augment our data with average age, percentage of male traffic controllers, and percentage of part-time workers (i.e., factors which are not taken into account by the Risk Index model).

Although the impact of automation is beyond the scope of our study, we will take account of this aspect in our regression models through a proxy variable, capturing automation levels. Real-time railway traffic control is a complex process which cannot be easily fully automated (Balfe et al., 2015). The progressive automation of the railway traffic systems is expected have a positive impact on safety and operational efficiency (Strategic Rail Research and Innovation Agenda, European Rail Research Advisory Council, 2014[3]). However, there are a number of issues which may raise concern, such as the impact of automation on situational awareness (see e.g. Lo et al., 2016). In order to control for effects of automation, we calculate the percentage of automatically signaled movements. This variable is defined as the ratio of automatically opened signals on the total number of signal openings (manual and automatic), and is multiplied by 100. The signal automation is based on a highly flexible system, where the traffic controller can decide to change the automation levels for each individual signal, for each individual train route, or for entire groups of trains and signals. In addition, all 11 control centres under study are also equipped with the so-called Automatic Route Setting system (automatically setting the routing of trains, or changing routes in order to minimize delay). The signal automation variable does not capture the route setting automation, but it does provide a reasonable proxy for the overall level of automation. Also, to control for local operational idiosyncrasies (e.g. in in track layout complexity, see Roets and Christiaens (2015), we add Traffic Control Centre fixed effect (dummy) variables to the disaggregated

[3]The European Rail Research Advisory Council (ERRAC) is a joint initiative of the European Commission and European railway industry stakeholders.

dataset. Finally, to control for time trends and seasonal effects, we complete the data with monthly dummy variables.

Tables 3.2 and 3.3 present the descriptive statistics of the 6 datasets. An in-depth data examination, enabled by the Business Intelligence tool, revealed issues with the work shifts during, before and after national or local railway strikes. In addition, several observations exhibited errors in the data, mainly during the transition from one month to the next. Therefore, 20 observations were deleted from the aggregated dataset, and 270 observations from the disaggregated dataset. This leads to a total number of 1,078 work shifts in the full week aggregated dataset, and 11,808 in the full week disaggregated dataset.

Table 3.2: Descriptive statistics for the aggregated dataset

	Full week				**Weekdays**		**Weekends**	
	Mean	St. dev.	Min	Max	Mean	St. dev.	Mean	St. dev.
Error rate	5.104	2.242	1.549	16.588	4.903	1.861	5.557	2.875
Avg. Risk Index	0.882	0.131	0.674	1.287	0.867	0.109	0.915	0.166
Avg. age	43.505	1.126	39.708	46.830	43.463	1.142	43.597	1.086
% Male	90.929	3.146	75.634	100.000	90.775	3.074	91.274	3.280
% Part-time	11.137	3.258	1.333	22.680	11.485	3.215	10.354	3.224
% Auto signaled	76.395	3.524	64.470	88.920	75.355	2.412	78.731	4.404
Observations	N= 1,078				N= 746		N= 332	

3.3.2 Regression analysis

In our real-life traffic control analysis, the aggregated datasets do not exhibit zero values in traffic control error rate, while the full week disaggregated dataset contains 11% zero values (with 8% for the weekday and 18% for the weekend dataset). We therefore apply Ordinary Least Squares (OLS) regression to the aggregated datasets, and adopt a Tobit regression model (Tobin, 1958) for left-censored data for the disaggregated datasets. Applying OLS regression on censored data produces biased and inconsistent parameter estimates. In addition, the Tobit results can further deepen empirical and managerial insight by revealing (decomposed)

Table 3.3: Descriptive statistics for the disaggregated dataset

	Full week				**Weekdays**		**Weekends**	
	Mean	St. dev.	Min	Max	Mean	St. dev.	Mean	St. dev.
Error rate	4.764	5.151	0.000	83.333	4.569	4.269	5.202	6.705
Avg. Risk Index	0.883	0.144	0.673	1.535	0.867	0.118	0.918	0.184
Avg. age	44.119	4.710	29.367	58.921	44.122	4.642	44.111	4.860
% Male	91.633	11.196	26.506	100.000	91.571	11.005	91.772	11.615
% Part-time	11.790	13.575	0.000	83.333	11.986	13.350	11.350	14.060
% Auto signaled	77.868	9.721	0.000	100.000	76.937	9.260	79.965	10.386
Observations	N= 11,808				N= 8,176		N= 3,632	

marginal effects.

The Tobit model has extensively proven its merits in econometric studies, and has found several fruitful applications in road safety research (Anastasopoulos et al., 2008; Debnath et al., 2014; Bin Islam and Hernandez, 2016). Despite its widespread use, the fatigue literature has not adopted Tobit regressions[4], and the approach has never been applied in conjunction with fatigue risk models. We briefly present the essentials of the Tobit model, and refer the interested reader to Breen (1996) and Wooldridge (2015) for a more detailed discussion.

The Tobit model relates the observed dependent variable y, the number of human errors per 1000 train movements, with an underlying latent (unobserved) variable y^*. For observation i this can be written as:

$$y_i = \left\{ \begin{array}{ll} y_i^* & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* \leqslant 0 \end{array} \right\} \tag{3.1}$$

In other words, the observed variable y equals the latent variable y^* if it is above zero, and is zero otherwise. The latent variable y^* is modeled as the classical linear model:

$$y_i = \mathbf{x}_i \boldsymbol{\beta} + u_i \tag{3.2}$$

[4]The Bennett and Passmore (1984) coal miner injury analysis being an early and notable exception.

Where $\mathbf{x}_i$ is the vector of independent variables, $\boldsymbol{\beta}$ the vector of estimable parameters, and u_i an independent and normally distributed error term, with zero mean and constant variance σ^2. The vector $\boldsymbol{\beta}$ is estimated through Maximum Likelihood Estimation.

In addition to the sign and statistical significance of the parameter estimates β_j, the effect of a change of the independent variable x_j on the outcome variable y can provide highly relevant information. This commonly-called 'marginal effect' is fairly simple to obtain for the latent variable $y*$ (and is equal to the parameter estimate β_j). However, due to the non-linearity of the Tobit model, it is more complex to estimate for the actual variable y. If x_j is a continuous variable, the marginal effect can be obtained by differential calculus. Formally, the marginal effect of x_j on $E(y \mid \mathbf{x})$, i.e. the expected value of y conditional on $\mathbf{x}$, can be obtained through the partial derivative of x_j. It can be shown that:

$$\frac{\partial E(y \mid \mathbf{x})}{\partial x_j} = \beta_j \Phi(z) \tag{3.3}$$

where $z = \frac{\mathbf{x}\boldsymbol{\beta}}{\sigma}$, and $\Phi(z)$ is the standard normal cumulative distribution function.

McDonald and Moffitt (1980) propose an insightful split of this marginal effect in two terms[5] :

$$\frac{\partial E(y \mid \mathbf{x})}{\partial x_j} = \left(\frac{\partial P(y > 0 \mid \mathbf{x})}{\partial x_j} \right) E(y \mid y > 0, \mathbf{x}) + P(y > 0 \mid \mathbf{x}) \left(\frac{\partial E(y \mid y > 0, \mathbf{x})}{\partial x_j} \right) \tag{3.4}$$

As such, the total change in y can be written in terms of (i) the change in the probability of being above zero, weighted by $E(y \mid y > 0, \mathbf{x})$, the expected value of y for those observations above zero; and (ii) the change in y along these observations above zero, weighted by the probability of being above zero $P(y > 0 \mid \mathbf{x})$. It can be shown that:

$$\frac{\partial P(y > 0 \mid \mathbf{x})}{\partial x_j} = \beta_j \frac{\phi(z)}{\sigma} \tag{3.5}$$

[5]Briefly explained, $E(y \mid \mathbf{x})$ can be written as the product $P(y > 0 \mid \mathbf{x}) E(y \mid y > 0, \mathbf{x})$; i.e. the probability of being above zero, multiplied by the expected value of the subpopulation above zero. Also, the derivative of a product of two functions $f(x)g(x)$ is given by $[f(x)g(x)]' = f(x)'g(x) + f(x)g(x)'$.

and

$$\frac{\partial E(y \mid y > 0, \mathbf{x})}{\partial x_j} = \beta_j \left[1 - z\frac{\phi(z)}{\Phi(z)} - \left(\frac{\phi(z)}{\Phi(z)}\right)^2 \right] \tag{3.6}$$

with $\phi(z)$ the standard normal density function.

In other words, in the case of our traffic controller error rates, the marginal effect of x_j on the expected error rate $E(y \mid \mathbf{x})$ can be decomposed in two intuitively appealing and useful 'subeffects' (Breen, 1996). First, the marginal effect of x_j on the *probability* of having an error rate above zero; and second, the marginal effect of an independent variable x_j on the *magnitude* of the error rate, given that errors do occur. If x_j is not a continuous but a binary variable, such as a day-of-week dummy variable, the effect of interest can be obtained by calculating the difference of values obtained with $x_j = 1$ and with $x_j = 0$, all other variables x_k $(k \neq j)$ being held constant (Wooldridge, 2015). To obtain an overall indication of the magnitude of the marginal effects, we calculate the mean marginal effects across all observations. The alternative, calculating the marginal effects of the average observation $\bar{\mathbf{x}}$, would imply unrealistic mean values $\bar{x_j}$ for the binary values. The pseudo R-squared of the Tobit regression model is obtained by calculating the the square of the correlation coefficient between the actual response values y and the expected values $E(y \mid \mathbf{x})$ (Wooldridge, 2015).

3.3.3 Out-of-sample cross-validation

In order to test the predictive accuracy of the regression models, we perform a 5-fold cross-validation procedure. K-fold cross-validation is a well-established data mining technique for gauging the performance of predictive models. It randomly partitions the original dataset into k equally sized subsets. One subset is held out for validation purposes, while the data from the remaining k-1 subsets is applied to estimate the regression parameters. With k = 5, 80% of the data is building the model, while 20% is held out for out-of-sample validation (by calculating the correlation between the observed error rates in the validation subset, and the error rates predicted by the estimated model). The process is repeated for each of the 5 subsets, and the average correlation coefficient of the out-of-sample predictive experiments is retained as measure of predictive performance. The advantage of the method is that all observations from the original dataset are used for estimation as well as validation.

3.4 Results and Discussion

3.4.1 Correlation analysis

Table 3.4 displays, for each of the six datasets, the Pearson correlation coefficients between the observed error rates and the Risk Index. All correlations in the table are significant at the 1 percent level. Given the real-world setting of our analysis, the observed correlations (0.583 - 0.678 for the aggregated dataset and 0.213 - 0.276 for the disaggregated dataset) perform well and are highly in line with correlation strengths found in previous field-based research (Fletcher and Dawson, 2001)[6]. Not unexpectedly, the empirically observed relationships are weaker in real-life than in (the carefully controlled settings of) laboratory-based validation studies. The breaking down of the strong relationships observed under laboratory conditions can at least partly be attributed to non-work-related factors, such as family and domestic responsibilities, which compete in real-life with the need for recovery sleep (see ibidem). As such, our correlation results validate the Risk Index in a workplace environment, more specifically in a railway traffic control setting.

Table 3.4: Correlations of the observed error rate with the Risk Index

	Full week	**Weekdays**	**Weekends**
Aggregated dataset	0.638	0.678	0.583
Disaggregated dataset	0.249	0.276	0.213

3.4.2 Regression analysis and cross-validation

The details of the 6 regression models are presented in table 3.5. As explained in the data and methodology section (see figure 1 in section 3.1), we focus on the full week Tobit regressions, and consider the aggregated and working week/weekend results as robustness checks. For confidentiality reasons, the regression estimates for the monthly and Traffic Control Centre fixed effect variables are not reported.

[6]Examining the correlations between (the two-step) predicted fatigue, (subjective) alertness levels and (objective) performance levels, correlation strengths in this study ranged from 0.10 to 0.25. When decomposing the relationships in function of additional parameters such as shift duration or time of day, the correlations are not consistently significant and exhibit a much wider range (from 0.00 to 0.60).

Table 3.5: Regression results

Variable	**OLS regression results (aggregated dataset)**			**Tobit regression results (disaggregated dataset)**		
	Full week	**Weekdays**	**Weekend**	**Full week**	**Weekdays**	**Weekend**
Constant	16.284***	16.128***	15.736**	18.407***	15.311***	24.194***
	(3.018)	(3.129)	(6.812)	(1.061)	(1.108)	(2.478)
Roster-based risk						
Average Risk Index	7.161***	8.659***	5.683***	5.138***	6.598***	3.164***
	(0.526)	(0.595)	(1.024)	(0.362)	(0.423)	(0.709)
Day-of-week effect						
Tuesday	-0.409**	-0.484***		-0.380**	-0.423***	
	(0.174)	(0.142)		(0.175)	(0.141)	
Wednesday	-0.470***	-0.584***		-0.412**	-0.506***	
	(0.179)	(0.148)		(0.178)	(0.144)	
Thursday	-0.505***	-0.629***		-0.344*	-0.456***	
	(0.179)	(0.149)		(0.179)	(0.144)	
Friday	-0.0005	-0.159		0.155	0.024	
	(0.180)	(0.151)		(0.178)	(0.145)	
Saturday	1.438***		1.256***	1.218***		1.146***
	(0.196)		(0.230)	(0.180)		(0.239)
Sunday	0.083			-0.067		
	(0.213)			(0.180)		
Individual characteristics						
Average age	-0.069	-0.051	-0.036	-0.025*	-0.028**	-0.023
	(0.046)	(0.046)	(0.107)	(0.014)	(0.013)	(0.032)
% Male	0.038**	0.020	0.057*	0.0001***	0.0001**	0.0002
	(0.015)	(0.015)	(0.034)	(0.00005)	0.00005)	(0.0001)
% Part-time	-0.045***	-0.039**	-0.040	-0.002	-0.004	0.005
	(0.016)	(0.016)	(0.037)	(0.004)	(0.004)	(0.011)
Operational characteristics						
% Automatically signaled	-0.229***	-0.230***	-0.249***	-0.232***	-0.197***	-0.298***
	(0.021)	(0.024)	(0.040)	(0.008)	(0.008)	(0.017)
N	1,078	746	332	11,808	8,176	3,632
Adjusted R2	0.556	0.577	0.542			
Pseudo R2				0.222	0.239	0.215

* p ¡ 0.1; ** p ¡ 0.05; *** p ¡ 0.01; standard errors between brackets.

Dummy variables for Traffic Control Centres and months not reported.

The OLS models explain a large proportion of the error rate variance (adjusted R-squared of 54-58%). For the Tobit regressions the pseudo R-squared varies between 21 and 24% (22% for the full week model). All of the 6 regression models indicate a positive and highly significant effect of the average team Risk Index on error rates, thus corroborating the correlation-based validation discussed above. The coefficient estimates of the other variables show largely consistent signs across the 6 regression results, but with sometimes diverging statistical significance. With regard to the variables age, gender, and part-time work, largely consistent signs are found, but no consistent statistical significance can be observed. As such, our results do not provide evidence of an effect of these variables on the error rate outcome. The operational variable controlling for the level of signaling automation is highly significant across all models, and reduces the error rate. As stated in the model specification section, this variable does not fully capture all levels of automation in the Traffic Control Centres. In addition, the complexity of real-time railway control is characterized by a high variability and uncertainty (Ferreira and Balfe, 2014). As such, possible detrimental effects of automation on human performance may be masked by averaging out the automation levels over the entire 8-hour work shift, and across the different workstations in the Traffic Control Centre.

Importantly, our regression results reveal consistently significant effects for several days of the week. Monday, the reference day for the full week as well as the weekday regression models, exhibits a significantly higher impact on error rate than Tuesday, Wednesday, and Thursday. Only on Saturdays a significantly higher effect can be observed. The effect on Saturdays is also significantly higher in the weekend regression models (compared to Sundays). The observed effects are largely in agreement with previous research examining day-of-week trends (Monk and Wagner, 1989; Brogmus, 2007; Wirtz et al., 2011).

The results of the 5-fold cross-validation tests can be found in table 3.6. It displays (i) the correlations between the observed error rates and the regression predicted error rates and (ii) the average correlations obtained by the cross-validation procedure. For all 6 models, the correlation strengths are clearly confirmed by the by the 5-fold cross-validation procedure. We can also see that the correlation strengths are substantially higher than the error rate correlations with the 'stand-alone' Risk Index (table 4). Correlation rises substantially to 0.751 - 0.782 for the aggregated dataset, and to a range of 0.465 – 0.489 for the disaggregated

dataset.

Table 3.6: Correlations of observed error rates with regression predictions

Aggregated dataset	**Full week**	**Weekdays**	**Weekends**
OLS model	0.751	0.767	0.782
OLS (5-fold cross-validation)	0.732	0.744	0.711
Disaggregated dataset	**Full week**	**Weekdays**	**Weekends**
Tobit model	0.471	0.489	0.465
Tobit (5-fold cross-validation)	0.464	0.482	0.452

3.4.3 Marginal effects

Table 3.7 reports the average marginal effects, for the Risk Index and the day-of-week variables. We specifically examine the decomposed marginal effects. As explained in the data and methodology section, the marginal effects of continuous variables (the Risk Index) provide a sense of the impact of a one-unit change on the error rate. For the discrete variables (the day-of-week variables), they measure the effect of switching from Mondays to another day in the week. The first column repeats the Tobit parameter estimate displayed in table 5. The second column presents the change in the probability of having at least one error, the third column the impact on the magnitude of the error rate, given that errors do occur. The marginal effects indicate that, on average, a unit increase of the Risk Index leads to a 27 percentage point increase in the probability of having at least one error, and an error rate increase of 3 errors per 1000 train movements. Or, expressed differently, a 0.1 increase of the Risk Index (which ranges from 0.673 to 1.535 in the considered dataset) leads to a 2.7 percentage point increase in the probability of having at least one error.

When zooming in on the impact of the day of the week, we observe that the probability of having at least one error is highest on Saturdays (a 5.92 percentage point higher probability compared to Mondays), and lowest on Tuesday, Wednesday and Thursday (around 2 percentage point decrease in probability). At first sight, this result might seem contradictory with the higher number of zero error occurrences we observed in our weekend dataset. However, the regression marginal effects are based on the ceteris paribus assumption that all other

factors (besides the day of the week) are held unchanged. Stated differently, on Saturdays, compared to Mondays, and all other circumstances being the same (such as automation levels or local idiosyncrasies), one can expect a higher probability of having at least one error.

Table 3.7: Average marginal effects for the full week Tobit model

	Parameter estimate	Change in error probability	Change in error magnitude
Average Risk Index	5.138	27.31%	3.070
Tuesday	-0.380	-2.08 %	-0.223
Wednesday	-0.412	-2.26 %	-0.241
Thursday	-0.344	-1.88 %	-0.202
Friday	0.155	ns	ns
Saturday	1.218	5.92 %	0.766
Sunday	-0.067	ns	ns

ns: non-significant Tobit parameter estimation

day-of-week effects are relative to Monday

3.5 Conclusions

In this paper, we examine the human factor component of railway traffic control safety, and more specifically fatigue risk and its link with human error. We evaluate the predictive validity of the Risk Index, a work schedule-based fatigue risk tool, and investigate the effect of additional risk factors (age, gender, part-time work, and day-of-week) on human error probability. By linking workforce and operational data, we are able to generate datasets which contain the necessary information at a highly disaggregate level (more than 11,000 work shifts are examined, including data on human error occurrence). Our approach can easily be applied to other fatigue risk models.

The empirical results validate the applied fatigue risk model in a railway traffic control setting. This extends previous research on the validity of two-step (work schedule-based) fatigue risk models in general, and the Risk Index in particular. As indicated by Dawson et al. (2011), there is no significant literature on the validity of fatigue risk models, in part because the field is relatively young. With specific regard to the validity of the applied Risk

Index, our results augment the previous web survey-based research by Greubel and Nachreiner (2013) to railway traffic control settings. As such, we provide the first validation study of the Risk Index in a real-life transportation environment, more particularly in the railway industry, where the model is widely used.

Our regression results do not indicate a significant effect of individual characteristics such as age, gender, or part-time work. However, consistently significant day-of-week effects are observed, during and surrounding the weekend. This result quantitatively supports the suggestion by Greubel and Nachreiner (2013) to reinforce the accuracy of the Risk Index by accounting for 'day-of-week' effects. Hence, our empirical research also contributes to the scarce but steadily growing fatigue risk literature on the day-of-week influences. Compared to Mondays, the regression predicted error rate is highest on Saturdays. All other things being equal, there is a 6 percentage point higher probability on Saturdays to make at least one error. In addition, Tuesdays, Wednesdays, and Thursdays show a 2 percentage point lower probability. Beyond its empirical value, this quantitative information is also highly relevant to transportation practitioners, and can be easily conveyed to non-statisticians across the company. As such, it can support management policy and communication with actionable language. More in general, the results suggest that safe work schedule design should also take into account the day of the week, and not exclusively rely on current fatigue model outputs. This conclusion is a clear demonstration of the current lack of fatigue risk models to account for non-biological fatigue factors, such as social or family activities (Dawson et al., 2011; Di Milia et al., 2011). As such, this paper not only adds to the under-researched area of railway traffic controller fatigue, but also responds to the fatigue risk modelling deficiencies reported in the literature.

In addition to the above-mentioned policy-relevant measures, our research also aims to actively bridge the gap between theory and practice. The data-driven and quantitative approach is perceived by the railway management as non-intrusive, flexible and easily implementable. The systematic approach also allows to move beyond the classical limitations of a one-shot exercise, and set up a permanent monitoring of fatigue risk levels and human errors. Importantly, empirical results are expressed in operationally intelligible metrics such as 'the number errors per 1000 train movements' or 'the probability of making at least one error'. Looking

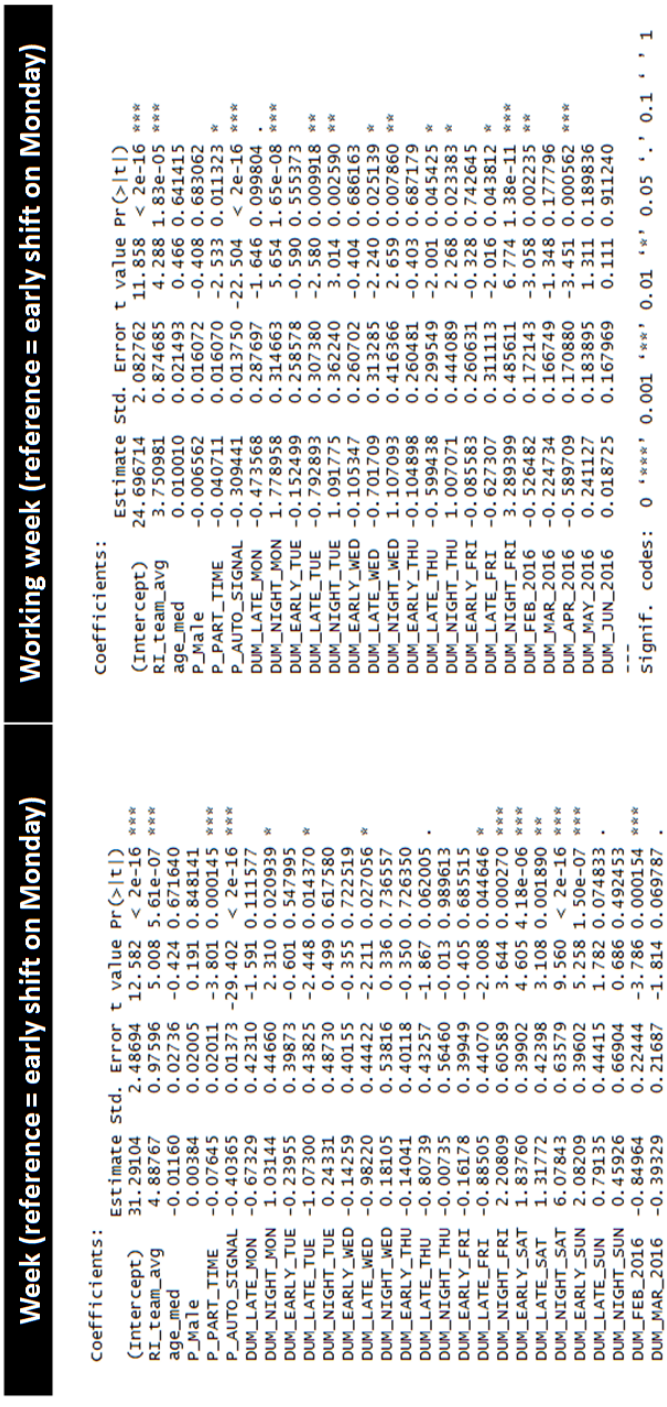
forward, our custom-developed Business Intelligence tool will be further developed to act as a decision support tool for senior management, and allow in-depth risk analysis by railway experts with extensive field experience. The objective is to quantitatively support and monitor strategic staffing decisions, not only concerning work schedule risk but also regarding staff efficiency levels. In the near future, the idea is to further explore the available databases to obtain human error data at individual instead of team level, and further improve on the data capturing automation levels.

To conclude, we believe that by systematically linking workforce and operational data, a wide range of safety-relevant topics can be explored, not only in railways but also in other sectors. With regard to railways, the ongoing digitization of traffic control systems in Europe (Wilson and Norris, 2005; Roets and Christiaens, 2015) provides an excellent and timely opportunity for this data-driven research. The systematic data link also allows to set up appropriate safety monitoring systems, which could convince potentially interested (railway) companies to step into the research project. Future research efforts could broaden the scope of investigation by also considering other human factors besides fatigue, such as staff skills, local knowledge of the infrastructure, or mental workload.

Acknowledgements

The authors would like to thank Infrabel for funding this research, and providing the necessary railway data and expertise. Special thanks goes to the management and experts of the Traffic Operations department, to the Methods department, as well as to the Performance Data division and the Business Analytics team. It is to be noted that the views expressed in this paper are those of the authors and do not necessarily reflect the opinions of Infrabel.

Appendix: interactions between day-of-week and shift type



Possible interactions between day-of-week and shift type can be examined by re-estimating the Risk Index at hourly level (instead of its conventional work shift level). Following Folkard et al. (2006), we calculated this hourly version of the Risk Index through cosinor regressions.

References

Åkerstedt, T. (2000). Consensus statement: fatigue and accidents in transport operations. *Journal of sleep research*, 9(4):395–395.

Anastasopoulos, P. C., Tarko, A. P., and Mannering, F. L. (2008). Tobit analysis of vehicle accident rates on interstate highways. *Accident Analysis & Prevention*, 40(2):768–775.

Anund, A., Fors, C., Kecklund, G., Leeuwen, W. v., and Åkerstedt, T. (2015). *Countermeasures for fatigue in transportation: a review of existing methods for drivers on road, rail, sea and in aviation*. Statens väg-och transportforskningsinstitut.

Balfe, N., Sharples, S., and Wilson, J. R. (2015). Impact of automation: measurement of performance, workload and behaviour in a complex control environment. *Applied ergonomics*, 47:52–64.

Bennett, J. D. and Passmore, D. L. (1984). Days lost from work due to injuries in us underground bituminous coal mines, 1975–1981. *Journal of occupational accidents*, 5(4):265–278.

Bin Islam, M. and Hernandez, S. (2016). Fatality rates for crashes involving heavy vehicles on highways: A random parameter tobit regression approach. *Journal of Transportation Safety & Security*, 8(3):247–265.

Breen, R. (1996). *Regression models: Censored, sample selected, or truncated data*, volume 111. Sage.

Brogmus, G. (2007). Day of the week lost time occupational injury trends in the us by gender and industry and their implications for work scheduling. *Ergonomics*, 50(3):446–474.

Civil Aviation Safety Authority (2014). Biomathematical fatigue models, guidance document. *Civil Aviation Safety Authority Australia*.

Cotrim, T., Carvalhais, J., Neto, C., Teles, J., Noriega, P., and Rebelo, F. (2017). Determinants of sleepiness at work among railway control workers. *Applied ergonomics*, 58:293–300.

Darwent, D., Dawson, D., Paterson, J. L., Roach, G. D., and Ferguson, S. A. (2015). Managing fatigue: It really is about sleep. *Accident Analysis & Prevention*, 82:20–26.

Dawson, D., Darwent, D., and Roach, G. D. (2017). How should a bio-mathematical model be used within a fatigue risk management system to determine whether or not a working time arrangement is safe? *Accident Analysis & Prevention*, 99:469–473.

Dawson, D., Noy, Y. I., Härmä, M., Åkerstedt, T., and Belenky, G. (2011). Modelling fatigue and the use of fatigue models in work settings. *Accident Analysis & Prevention*, 43(2):549–564.

Dean, D. A., Fletcher, A., Hursh, S. R., and Klerman, E. B. (2007). Developing mathematical models of neurobehavioral performance for the "real world". *Journal of Biological Rhythms*, 22(3):246–258.

Debnath, A. K., Blackman, R., and Haworth, N. (2014). A tobit model for analyzing speed limit compliance in work zones. *Safety science*, 70:367–377.

Di Milia, L., Smolensky, M. H., Costa, G., Howarth, H. D., Ohayon, M. M., and Philip, P. (2011). Demographic factors, fatigue, and driving accidents: An examination of the published literature. *Accident Analysis & Prevention*, 43(2):516–532.

Dorrian, J., Baulk, S. D., and Dawson, D. (2011). Work hours, workload, sleep and fatigue in australian rail industry employees. *Applied ergonomics*, 42(2):202–209.

European Rail Research Advisory Council (2014). Strategic rail research and innovation agenda.

Ferreira, P. N. and Balfe, N. (2014). The contribution of automation to resilience in rail traffic control. In *International Conference on Engineering Psychology and Cognitive Ergonomics*, pages 458–469. Springer.

Fletcher, A. and Dawson, D. (2001). Field-based validations of a work-related fatigue model based on hours of work. *Transportation research part F: traffic psychology and behaviour*, 4(1):75–88.

Folkard, S. and Lombardi, D. A. (2004). Toward a "risk index" to assess work schedules. *Chronobiology international*, 21(6):1063–1072.

Folkard, S. and Lombardi, D. A. (2006). Modeling the impact of the components of long work hours on injuries and "accidents". *American journal of industrial medicine*, 49(11):953–963.

Folkard, S., Lombardi, D. A., and Spencer, M. B. (2006). Estimating the circadian rhythm in the risk of occupational injuries and accidents. *Chronobiology international*, 23(6):1181–1192.

Folkard, S., Robertson, K. A., and Spencer, M. B. (2007). A fatigue/risk index to assess work schedules. *Somnologie-Schlafforschung und Schlafmedizin*, 11(3):177–185.

French, J. and Neville, K. J. (2012). Avoiding the impact of fatigue on human effectiveness. In *The handbook of operator fatigue*. Ashgate Publishing Ltd.

Friedl, K. E., Mallis, M. M., Ahlers, S. T., Popkin, S. M., and Larkin, W. (2004). Research requirements for operational decision-making using models of fatigue and performance. *Aviation, space, and environmental medicine*, 75(3):A192–A199.

Gander, P., Hartley, L., Powell, D., Cabon, P., Hitchcock, E., Mills, A., and Popkin, S. (2011). Fatigue risk management: Organizational factors at the regulatory and industry/company level. *Accident Analysis & Prevention*, 43(2):573–590.

Gertler, J., DiFiore, A., and Raslear, T. (2013). Fatigue status of the us railroad industry. Technical Report DOT/FRA/ORD-13/06, US Department of Transportation, Federal Railroad Administration, Washington DC.

Gertler, J. and Viale, A. (2007). Work schedules and sleep patterns of railroad dispatchers. Technical Report DOT/FRA/ORD-07/11, US Department of Transportation, Federal Railroad Administration, Washington DC.

Greubel, J. and Nachreiner, F. (2013). The validity of the risk index for comparing the accident risk associated with different work schedules. *Accident Analysis & Prevention*, 50:1090–1095.

Härmä, M., Sallinen, M., Ranta, R., Mutanen, P., and Müller, K. (2002). The effect of an irregular shift system on sleepiness at work in train drivers and railway traffic controllers. *Journal of sleep research*, 11(2):141–151.

Lerman, S. E., Eskin, E., Flower, D. J., George, E. C., Gerson, B., Hartenbaum, N., Hursh, S. R., Moore-Ede, M., et al. (2012). Fatigue risk management in the workplace. *Journal of Occupational and Environmental Medicine*, 54(2):231–258.

Lo, J. C., Sehic, E., Brookhuis, K. A., and Meijer, S. A. (2016). Explicit or implicit situation awareness? measuring the situation awareness of train traffic controllers. *Transportation research part F: traffic psychology and behaviour*, 43:325–338.

Mallis, M. M., Mejdal, S., Nguyen, T. T., and Dinges, D. F. (2004). Summary of the key features of seven biomathematical models of human fatigue and performance. *Aviation, space, and environmental medicine*, 75(3):A4–A14.

Marucci-Wellman, H. R., Lombardi, D. A., and Willetts, J. L. (2016). Working multiple jobs over a day or a week: Short-term effects on sleep duration. *Chronobiology international*, 33(6):630–649.

McDonald, J. F. and Moffitt, R. A. (1980). The uses of tobit analysis. *The review of economics and statistics*, pages 318–321.

Monk, T. H. and Wagner, J. A. (1989). Social factors can outweigh biological ones in determining night shift safety. *Human Factors*.

National Transportation Safety Board (2017). Most wanted list. https://www.ntsb.gov/mostwanted. Accessed: 2017-02-03.

Pachl, J. (2009). *Railway operation and control*. VTD Rail Publishing, Mountlake Terrace (USA).

Popkin, S., Gertler, J., and Reinach, S. (2001). A preliminary examination of railroad dispatcher workload, stress, and fatigue. Technical report.

Rail Safety and Standards Board (2015). Fatigue and its contribution to railway incidents.

Raslear, T. G., Gertler, J., and DiFiore, A. (2013). Work schedules, sleep, fatigue, and accidents in the us railroad industry. *Fatigue: Biomedicine, Health & Behavior*, 1(1-2):99–115.

Reason, J. (1990). *Human error*. Cambridge university press.

Roets, B. and Christiaens, J. (2015). Evaluation of railway traffic control efficiency and its determinants. *European Journal of Transport and Infrastructure Research*, 15(4):396–418.

Sallinen, M., Härmä, M., Mutanen, P., Ranta, R., Virkkala, J., and Müller, K. (2005). Sleepiness in various shift combinations of irregular shift systems. *Industrial health*, 43(1):114–122.

Spencer, M., Robertson, K., and Folkard, S. (2006). The development of a fatigue/risk index for shiftworkers. *Health and Safety Executive Report*, 446.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica: journal of the Econometric Society*, pages 24–36.

Wilson, J. R. and Norris, B. J. (2005). Rail human factors: Past, present and future. *Applied ergonomics*, 36(6):649–660.

Wirtz, A., Nachreiner, F., and Rolfes, K. (2011). Working on sundays–effects on safety, health, and work-life balance. *Chronobiology international*, 28(4):361–370.

Wooldridge, J. M. (2015). *Introductory econometrics: A modern approach*. Nelson Education.

Chapter 4

MULTI-OUTPUT EFFICIENCY AND HUMAN ERROR

Abstract

Transportation service providers are under increasing pressure to raise cost efficiency without sacrificing safety. This paper advocates a multi-output cost efficiency framework to monitor staffing efficiency, and shows the relation between multi-output efficiency and human error. To realistically model input-output relations, we propose a Data Envelopment Analysis (DEA)-based framework with proportional cost allocation restrictions. Our analysis at an hourly rate of Belgian computerized railway traffic control centres shows that human error relates to both output-specific efficiency components and binding cost allocation restrictions. Further, we confirm that sufficient time is reserved for safety operations in the railway traffic control centres under study, but we reveal staff schedule inefficiencies.

Keywords: Data Envelopment Analysis; Human error; Output-specific efficiency; Input-output allocations; Railways

4.1 Introduction

Transportation efficiency and safety are two major and interlinked policy objectives, spurring European railways to go through a profound change process. Since 1991, European directives have gradually unbundled the railway system into national '*infrastruc-*

The work in this chapter is co-authored by Johan Christiaens and Marijn Verschelde.

ture managers', and several competing railway undertakings. With the European Directive 2012/34/EU (2012), the European Union is pursuing the development of a competitive Single European Railway Area. It considers railway infrastructure as a natural monopoly, and as such urges the infrastructure managers to reduce costs. At the same time, railway safety levels should be maintained and even improved where practicable European Directive 2016/798/EU (2016).

In support of these challenges, we present a performance measurement tool for railway traffic control, a core railway infrastructure activity which leans heavily on efficiency and safety to improve performance. To model the input-output relations, we advocate a Data Envelopment Analysis (DEA)-based framework with proportional cost allocation restrictions. The framework pinpoints staff scheduling inefficiencies and reveals temporal effects. At the same time, our exceptionally disaggregated DEA analysis allows us to show how human error relates to multi-output efficiency and binding cost allocation restrictions. Measured staffing efficiency usually relates to heterogeneity in the way the outputs are produced by individuals (Leibenstein, 1966, 1973). In our setting, railway traffic control staff could - under conditions of work underload or overload - choose to produce their given tasks in a way or pace that correlates with more attention failure. In turn, attention deficits might indirectly impact safety levels. As such, by relating components of multi-output efficiency to human error (our proxy for attention failure), we quantitatively support decision makers in focusing on key efficiency parameters, while keeping safety in the spotlight. This demonstrates the usefulness of multi-output efficiency models in providing insights that go beyond mere productive efficiency considerations.

With this paper we grasp the opportunities offered by the continuing digitization of railway traffic control, and present data-driven empirical research on staffing efficiency that includes concerns about safety and human error. Using unique intra-company data from Infrabel, the Belgian railway infrastructure manager, we develop a DEA-based model to evaluate computerized traffic control. DEA is a well-established tool for measuring the efficiency of Decision Making Units (DMUs), which convert multiple inputs into multiple outputs (Charnes et al., 1978). Relying on Linear Programming techniques, the DEA methodology has sparked a large number of empirical efficiency studies in the past decades (for an overview, see the col-

lections of e.g. Fried et al., 2008 and Cooper et al., 2011). The distinguishing feature of the methodology is its non-parametric nature, implying that no a priori (typically unverifiable) functional form specifications are imposed on the production technology. As such, DEA is a data-oriented managerial decision-support tool.

To capture the traffic control process in a realistic way, we customize and extend the Cherchye et al. (2013) approach to open the black box of efficiency measurement, by formally including a priori information on the allocation of inputs to outputs. Our method parallels the approach proposed by Cook and Zhu (2011), who recognized that the relative importance of inputs can be output-specific. Traditionally, weight restrictions on DEA input multipliers impact the entire output bundle of the DMU. Cook and Zhu (2011) introduced output-specific weight restrictions, and encouraged further research along these lines. Our multi-output cost efficiency framework is based on cost allocation restrictions and has the distinguishing feature that it defines any type of input. This not only allows to tailor the input-output relations to the traffic control process at hand, but also to a wide range of other real-world settings. In addition, our restrictions are expressed in a proportional form, and as such have a straightforward, natural interpretation. Proportionally defined bounds are particularly appealing when a priori expert judgment needs to be translated into DEA restrictions (Sarrico and Dyson, 2004).

In addition to the developed multi-output cost efficiency framework, we further contribute to the efficiency literature in two respects. First, by calculating efficiency at an hourly resolution, we assess the around-the-clock efficiency of work shift schedules, and reveal within-shift patterns of inefficiency. As such we offer, to the best of our knowledge, an exceptionally disaggregated application of DEA. We have not identified any study in the service sector examining efficiency levels at hourly, daily, or even weekly basis. Notable exceptions of highly disaggregated efficiency studies can however be found in manufacturing (Hoopes and Triantis, 2001; Jain et al., 2011) and the fishing industry (Vázquez-Rowe and Tyedmers, 2013; Oliveira et al., 2014). Our traffic control application is developed in close collaboration with railway experts from Infrabel. Empirical results are based on a unique and rich data set of 11 computerized traffic control centres, covering every single hour of the entire year 2015. We support our framework with a custom-built Business Intelligence tool. Deployed at Infrabel's

central decision-making level, the Business Intelligence application actively bridges the gap between researchers and practitioners, and provides a real-life test case for the methodology.

Our empirical efficiency results provide quantitative insights in hour-of-day and day-of-week effects. By virtue of the applied multi-output methodology, we can further disentangle the efficiency scores and pinpoint the operational reasons underlying the observed efficiency patterns. We find that the output safety is hardly ever accompanied by a binding input cost allocation constraint. This means that we confirm that sufficient time is reserved for safety procedures. However, existing staff schedules are shown to not match the hour-of-day and day-of-week heterogeneity in railway traffic. Efficiency gains are expected by introducing a more flexible staff scheduling approach, tailored to the operational realities (e.g. by scheduling less but overlapping work shifts).

Second, by linking this exceptionally disaggregated information on multi-output efficiency with equally disaggregated data on human error, we pinpoint an empirical link between the components of multi-output efficiency and human error. The human errors considered in our study consist of relatively frequent but non safety-critical task errors, detected by the computerized traffic control system and subsequently archived for analytical purposes. Categorizable as attention failures (Reason, 1990), the human errors can indirectly impact safety in railway transportation. We relate the registered errors with the obtained multi-output efficiency results in a probit regression analysis. Fully leveraging the insights provided by the multi-output cost efficiency framework, we reveal a relation with the efficiency of production tasks with a highly variable work load, and with binding cost allocation restrictions. All results are robust for controlling for the number of movements, automation, and hour-of-day, week-day, month and traffic control centre fixed effects. In sum, the proposed framework and the exceptionally disaggregated data allow management to iteratively tackle efficiency issues and gradually move towards optimized hourly staffing levels. At the same time, it monitors safety operations and reveals relationships with human error.

The remainder of this paper is structured as follows. Section 2 presents the multi-output methodology and its cost allocation restricting extension. Section 3 describes the railway traffic control production process and the related human errors, and presents the data for the implementation of the framework. Section 4 reports and discusses the empirical results.

Conclusions are set out in the final section.

4.2 Multi-output efficiency framework

4.2.1 Methodological background

Closely related to our multi-output efficiency framework is the increasingly established Network DEA literature that deals with including information on interrelated sub-processes into a DEA analysis (see Färe and Grosskopf, 1996; Cook et al., 1998, 2000; Färe and Grosskopf, 2000; Fare et al., 2007; Cook and Zhu, 2014; and the reviews of e.g. Cook and Seiford, 2009; Castelli et al., 2010). This literature includes multi-output processes with both output-specific and shared resources (Beasley, 1995; Cook et al., 2000; Cook and Hababou, 2001). Shared inputs are modelled as having an unknown allocation over outputs. Based on axioms, Cherchye et al. (2013, 2014) define output-specific technology sets and introduce joint inputs, which are non-exclusively and non-rivalry used to produce the different outputs (see Lozano (2015) for the introduction of joint inputs in Network DEA). As joint inputs imply no allocation of the inputs over the outputs, Cherchye et al. (2013, 2014) propose to allocate output-specific prices of joint inputs, paralleling the idea Lindahl pricing, which is associated with Pareto-efficient public goods provision. In our framework, we develop and implement an unelaborated suggestion by Cherchye et al. (2013) to restrict the output-specific allocations. We formally extend their methodology by introducing proportional cost allocation restrictions, and show that the restrictions define any input type, including among others output-specific, shared, and joint inputs.

Appropriately restricting weights in DEA improves the discriminatory power of the inefficiency estimation and can avoid the possibility of unwanted specialisation (e.g., a zero weight for outputs related to safety). However, restricting multiplier weights often implies naturally undesirable weighting schemes (Cook and Seiford, 2009). Weight restrictions in multiplier models may also lead to infeasibilities (Allen et al., 1997) or may violate basic production assumptions by inducing free or unlimited production (Podinovski and Bouzdine-Chameeva, 2013, 2015; Podinovski, 2016). Further, the economic interpretation of multiplier weight restrictions is less straightforward than restrictions of the primal model (i.e., cost efficiency, revenue efficiency or profit efficiency). Cook and Zhu (2011) allow for output-specific input-

assurance regions. We introduce cost allocation restrictions that can be output-specific, and express the restrictions in a proportional form. By considering proportional cost share allocations rather than multiplier weighting schemes, we avoid the pitfalls related to multiplier weight restrictions and we obtain restrictions with a straightforward, natural interpretation.

4.2.2 Notational preliminaries

We start from a data set with T DMUs, which produce M outputs (denoted by a vector $\mathbf{y} \in \mathbb{R}_+^M$), using N inputs (denoted by a vector $\mathbf{X} \in \mathbb{R}_+^N$). Input prices are defined by $\mathbf{P} \in \mathbb{R}_+^N$. Formally, the full data set can be summarized as

$$S = \{\mathbf{y}_t, \mathbf{X}_t, \mathbf{P}_t | t = 1, ..., T\}. \tag{4.1}$$

A cornerstone of the multi-output approach is the concept of output-specific Input Requirement Sets. For each output, this set characterizes the output-specific production technology by containing all possible combinations of inputs that can produce the considered output quantity. More formally, the output-specific Input Requirement Set $I^m(y^m)$ for output m $(1 \leq m \leq M)$, consuming inputs $\mathbf{X}^m$, can be expressed as

$$I^m(y^m) = \left\{ \mathbf{X^m} |\ \mathbf{X^m} \in \mathbb{R}_+^N \text{ can produce } y^m \right\}. \tag{4.2}$$

The Input Requirement Sets for each output are assumed to be nested within each other:

$$y^m \geqslant y^{m*} \Rightarrow I^m(y^m) \subseteq I^m(y^{m*}). \tag{4.3}$$

This implies free disposability of outputs, i.e. we assume it is always the case that the DMUs can produce less output with the same input levels.

4.2.3 Modelling the production process

As developed in Cherchye et al. (2013, 2015), we account for inputs which can be of the output-specific, joint, or sub-joint type. See Cherchye et al. (2013, 2015) for a consideration of joint inputs, and Salerian and Chan (2005), Despić et al. (2007), and Cherchye et al. (2015) for the introduction of sub-joint inputs in DEA. We extend the multi-output efficiency framework by equally considering shared inputs (as introduced by Beasley, 1995; Cook et al., 2000; Cook and Hababou, 2001).

- *output-specific* inputs can be directly allocated to individual outputs and as such benefit the production process of the outputs involved. The allocation to each output is perfectly known. An example is an input '*employees*' which can be perfectly allocated to the production of specific outputs.

- *shared* inputs can also be allocated to individual outputs, but the allocation to the outputs is unobserved or not perfectly known. An example is an input '*maintenance staff*' of which only estimated but realistic allocation ranges to specific outputs are known.

- *joint* inputs cannot be allocated to specific outputs, but simultaneously assist in the production process of all inputs in a non-exclusive and non-rival way. In that sense, joint inputs act as '*public goods*' and as such account for economies of scope in the production process. An example is the input of a CEO.

- *sub-joint* inputs also act as public goods, but only for a subset of outputs. An example is an input '*quality control staff*', which is only used in the production of a limited number of outputs.

The basis for our cost efficiency for DMU t we elaborate below, is the cost share allocation to the M outputs. In particular, we fraction the input price $\mathbf{P}_t$ over the different outputs using so called '*output-specific prices*' $\mathbf{P}_t^m$. It are these output-specific prices that make the M output technologies interdependent. We can use a shadow price interpretation if the output-specific prices $\mathbf{P}_t^m$ and/or the aggregate prices $\mathbf{P}_t$ are unobserved. Formally, output-specific prices are any vectors $\mathbf{P}^m \in \mathbb{R}_+^N$ $(m \in 1, ..., M)$ that satisfy :

$$\sum_{m=1}^{M} \mathbf{P}^m = \mathbf{P}. \tag{4.4}$$

Although the mathematical modelling of the cost allocation mechanism through the output-specific prices is identical for all types of input, the interpretation is different. For shared inputs, the allocation of an input n to an output m can be represented by a fraction $A_n^m \in [0, 1]$. The cost share allocated to output m is then $(\mathbf{P})_n.\big(A_n^m.(\mathbf{X})_n\big)$, with $(\mathbf{X})_n$ representing input n and $(\mathbf{P})_n$ its price (which is independent of the output m). A_n^m allocates cost shares to the outputs through input allocation, and has the property $\sum_{m=1}^{M} A_t^m = 1$. In case of

an output-specific input, the fraction A_n^m is perfectly observed for all m. In line with the procedure proposed in Cook et al. (2000), we can rewrite the cost share as $\left(A_n^m.(\mathbf{P})_n\right).(\mathbf{X})_n$ and, by subsequently replacing $\left(A_n^m.(\mathbf{P})_n\right)$ with the output-specific price $(\mathbf{P}^m)_n$, as $(\mathbf{P}^m)_n.(\mathbf{X})_n$. The procedure is applied in order to preserve linearity in the mathematical models, see Cherchye et al. (2013) and Cook et al. (2000) for a more extensive discussion. The common shadow price $(\mathbf{P})_n$ for all outputs m reflects the assumption that shared inputs can be re-allocated over the outputs, in order to optimize efficiency (Cherchye et al., 2017).

For joint and sub-joint inputs, rather than allocating the inputs itself, Cherchye et al. (2013, 2014) propose to allocate their output-specific prices in a multi-output cost minimizing way (implying a cooperative perspective). This is economically meaningful as the allocated prices $\mathbf{P}^m$ have a Lindahl price interpretation. The respective output-specific prices correspond to the marginal production of the respective output (expressed in monetary terms) that follows from an additional unit of the joint inputs. The price allocation of joint inputs differs from allocating fixed costs of non-allocatable inputs as discussed in Cook and Kress (1999), as the latter focuses on the optimal allocation of costs related to fixed inputs, and thus abstracts from joint inputs that are under the discretion of the DMU.

In order to shape the input-output relationships with a priori knowledge of the production process, we impose proportional price restrictions on the ranges of the output-specific prices $\mathbf{P}^m$:

$$(\mathbf{L}^m)_n \leqslant \frac{(\mathbf{P}^m)_n}{(\mathbf{P})_n} \leqslant (\mathbf{U}^m)_n. \tag{4.5}$$

Here, $(\mathbf{L}^m)_n$ and $(\mathbf{U}^m)_n$ are elements of output-specific lower and upper bound vectors $\mathbf{L}^m$ and $\mathbf{U}^m$, which define the ranges for the ratios of output-specific prices to aggregate prices, with $\mathbf{L}^m \in \mathbb{R}_+^N$, $\mathbf{U}^m \in \mathbb{R}_+^N$, $\mathbf{0} \in \mathbb{R}_+^N$ and $\mathbf{1} \in \mathbb{R}_+^N$. As a direct consequence of the output-specific price definition (see formula 4.4), all elements $(\mathbf{L}^m)_n$ and $(\mathbf{U}^m)_n$ have dimensionless values between 0 and 1, and therefore $\mathbf{0} \leqslant \mathbf{L}^m \leqslant \mathbf{U}^m \leqslant \mathbf{1}$. Similar to the virtual weight restrictions developed by Wong and Beasley (1990), the nature of the proportional and dimensionless lower and upper bounds $\mathbf{L}^m$ and $\mathbf{U}^m$ facilitates the interaction between the empirical analyst and the business experts.

Through a judicious choice of the proportional upper and lower bounds for the output-specific (shadow) prices, we can now tailor the internal structure of the DMUs to the specific

production process under evaluation. As a practical illustration of this point, we consider the cost share allocations of a fictitious production process with 3 outputs, 2 output-specific inputs and 2 shared inputs. We bundle the upper and lower bound vectors $\mathbf{L}^m$ and $\mathbf{U}^m$ into a single input-output matrix $\in \mathbb{R}_+^{4\times 6}$, containing dimensionless elements $\in [0,1]$:

$$\begin{array}{c} \\ X_1 \\ X_2 \\ X_3 \\ X_4 \end{array} \begin{array}{c} \begin{array}{cc|cc|cc} \mathbf{L}^1 & \mathbf{U}^1 & \mathbf{L}^2 & \mathbf{U}^2 & \mathbf{L}^3 & \mathbf{U}^3 \end{array} \\ \left(\begin{array}{cc|cc|cc} 0.1 & 0.1 & 0.2 & 0.2 & 0.7 & 0.7 \\ 0 & 0 & 0.4 & 0.4 & 0.6 & 0.6 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0.2 & 0.4 & 0.3 & 0.6 & 0.2 & 0.5 \end{array}\right) \end{array} \begin{array}{l} \\ \textit{output-specific input} \\ \textit{output-specific input} \\ \textit{shared input with unobserved allocation} \\ \textit{shared input with allocation ranges.} \end{array}$$

In the presence of an additionally observed joint and sub-joint input, the input-output matrix associated with the production process can be augmented with the following matrix $\in \mathbb{R}_+^{2\times 6}$:

$$\begin{array}{c} \\ X_5 \\ X_6 \end{array} \begin{array}{c} \begin{array}{cc|cc|cc} \mathbf{L}^1 & \mathbf{U}^1 & \mathbf{L}^2 & \mathbf{U}^2 & \mathbf{L}^3 & \mathbf{U}^3 \end{array} \\ \left(\begin{array}{cc|cc|cc} 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \end{array}\right) \end{array} \begin{array}{l} \\ \textit{joint input} \\ \textit{sub-joint input.} \end{array}$$

Importantly, the overview provided by the bundling of the transparent allocation ranges in a single matrix $\in \mathbb{R}_+^{N\times 2M}$ strongly streamlines the interaction with business experts. The juxtapositioning of the ranges enables a global assessment of the feasibility, redundancy and operational logic of the allocations, not only for each individual input and output, but also for the production model as a whole. In addition, as we discuss in our railway traffic control application, the cost allocation bounds reveal information on human error.

4.2.4 Multi-output cost efficiency

For each output m and under the assumption of nested Input Requirement Sets, we benchmark the unit under evaluation DMU t against every output-dominating DMU s of data set S, i.e. against all DMUs that produce at least the output y_t^m, or, more formally, each DMU $s \in D_t^m = \{s|y_s^m \geqslant y_t^m\}$. For a given specification of output-specific prices $\mathbf{P}_t^m$, the minimal cost c_t^m for producing output m can then be defined as

$$c_t^m(\mathbf{P}_t^m) = \min_{s\in D_t^m} \left((\mathbf{P}_t^m)'\mathbf{X}_s\right). \tag{4.6}$$

The multi-output cost efficiency CE_t of DMU t is then the ratio between minimal costs to produce the M outputs and the actual costs:

$$CE_t(\mathbf{P}_t^1, ..., \mathbf{P}_t^M) = \frac{\sum_{m=1}^{M} c_t^m(\mathbf{P}_t^m)}{\mathbf{P}'\mathbf{X}_t}. \tag{4.7}$$

We have $0 \leq CE_t \leq 1$ with lower (higher) values indicating less (more) cost efficiency. Clearly, given the output-specific prices, multi-output efficient production implies production at the minimal cost for each output m ($m \in 1, ..., M$).

As discussed, output-specific prices $\mathbf{P}_t^m$ can be unobserved. We endogenise the choice of $\mathbf{P}_t^m$ by selecting the most favourable output-specific prices $\mathbf{P}_t^m$ in terms of overall cost efficiency. As such, we maximize CE_t over output-specific prices $\mathbf{P}_t^m$ ($1 \leq m \leq M$) and our empirical measure of cost efficiency is defined by

$$CE_t = \max_{(\mathbf{P}_t^1, ..., \mathbf{P}_t^M)} CE_t(\mathbf{P}_t^1, ..., \mathbf{P}_t^M). \tag{4.8}$$

4.2.5 Practical implementation

In case the input prices $\mathbf{P}_t$ are unobserved, the cost efficiency $\widehat{CE_t}$ of a DMU t can be estimated by solving the LP problem in (4.9). The multi-output cost efficiency measure has a direct dual interpretation as as an input-oriented Debreu-Farrell technical efficiency with output-specific input requirement sets. The detailed dual interpretation can be found in Cherchye et al. (2013, 2015).

The first three constraints of the following model implement the Cherchye et al. (2013) multi-output cost efficiency methodology. We augment the model by adding a formal structure

of proportional cost allocation restrictions.

$$
\begin{aligned}
\widehat{CE_t} = \underset{\substack{c_t^m \geqslant 0, \mathbf{P}_t \in \mathbb{R}_+^N \\ \mathbf{P}_t^m \in \mathbb{R}_+^N}}{\text{maximize}} \quad & \sum_{m=1}^{M} c_t^m \\
\text{subject to} \quad & \forall m : c_t^m \leqslant (\mathbf{P}_t^m)' \mathbf{X}_s, \quad \forall s \in D_t^m \qquad (C-1) \\
& \sum_{m=1}^{M} \mathbf{P}_t^m = \mathbf{P}_t \qquad (C-2) \\
& \mathbf{P}_t' \mathbf{X}_t = 1 \qquad (C-3) \\
& \mathbf{P}_t' \mathbf{R} \geqslant \mathbf{0} \qquad (C-4) \\
& \forall m : \mathbf{L}^m \leqslant \mathbf{P}_t^m \odot (\mathbf{P}_t)^{\odot(-1)} \leqslant \mathbf{U}^m. \quad (C-5)
\end{aligned}
\tag{4.9}
$$

The LP model maximizes the overall cost efficiency $\sum_{m=1}^{M} c_t^m$ of DMU t through optimizing values of the shadow prices $\mathbf{P}_t$ and $\mathbf{P}_t^m$. This '*most favourable pricing*' for evaluating the DMU t against other units, is similar to selecting the most favourable multiplier weights in a traditional DEA model (Cherchye et al., 2013). For each output m, constraint $C-1$ benchmarks the DMU under evaluation against every output-dominating DMU $s \in D_t^m$. As the LP objective function maximizes $\sum_{m=1}^{M} c_t^m$, the constraint assures that optimal shadow cost c_t^m does not exceed the shadow cost for DMU s in producing output m ($1 \leq m \leq M$), i.e $(\mathbf{P}_t^m)' \mathbf{X}_s$. Constraint $C-2$ incorporates the output-specific shadow prices definition. The normalization constraint $C-3$ (similar to the '*Charnes-Cooper transformation*' applied in Charnes et al., 1978) assures an efficiency value $\widehat{CE_t}$ between 0 and 1.

In constraint $C-4$, we restrict flexibility at the level of the aggregate shadow prices $\mathbf{P}_t$, by adding price restrictions which reflect realistic production trade-offs between the inputs (Podinovski, 2004). Following the general weight restriction formulation proposed by Halme and Korhonen (2000) and Joro and Korhonen (2015), we express the r shadow price restrictions in matrix form, with $\mathbf{R} \in \mathbb{R}^{N \times r}$ and $\mathbf{0} \in \mathbb{R}^r$ (row vector). This allows to define relative price restrictions which include, amongst others, the Assurance Region type I models developed by Thompson et al. (1986, 1990).

Constraint $C-5$ implements proportional cost allocation restrictions through output-specific shadow prices $\mathbf{P}_t^m$, with $\mathbf{L}^m \in \mathbb{R}_+^N$ and $\mathbf{U}^m \in \mathbb{R}_+^N$. For each input n and output m, it applies the lower and upper bounds $(\mathbf{L}^m)_n$ and $(\mathbf{U}^m)_n$ for the price fraction $\frac{(\mathbf{P}_t^m)_n}{(\mathbf{P}_t)_n}$ that is used to allocate cost shares. $(\mathbf{P}_t)^{\odot(-1)}$ represents the Hadamard inverse of $\mathbf{P}_t$, i.e. $((\mathbf{P}_t)^{\odot(-1)})_n = \frac{1}{(\mathbf{P}_t)_n}$.

Unique is that the input types are not explicitly modelled in the mathematical LP formulation, but embodied in the accompanying upper and lower bound vectors $\mathbf{L}^m$ and $\mathbf{U}^m$. In consequence, by merely changing the values of the allocation bounds, the empirical analyst can easily and flexibly capture the structure of a wide range of real-world production processes.

4.2.6 Efficiency decomposition

A highly attractive property of the multi-output efficiency model is its inherent capability to decompose the obtained efficiency value $\widehat{CE_t}$ of DMU t in output-specific efficiency values and corresponding weights. Given formulation (4.9) and given its normalization constraint $C-3$ ($\mathbf{P}_t'\mathbf{X}_t = 1$) this decomposition can be written as:

$$\widehat{CE_t} = \sum_{m=1}^{M} \widehat{c_t^m} = \sum_{m=1}^{M} \frac{(\widehat{\mathbf{P}_t^m})'\mathbf{X}_t}{\widehat{\mathbf{P}_t}'\mathbf{X}_t} \cdot \frac{\widehat{c_t^m}}{(\widehat{\mathbf{P}_t^m})'\mathbf{X}_t} = \sum_{m=1}^{M} \widehat{w_t^m} . \widehat{CE_t^m}. \tag{4.10}$$

The interpretation of $\widehat{w_t^m}$ and $\widehat{CE_t^m}$ is straightforward. For the given prices $\widehat{\mathbf{P}_t}$ and $\widehat{\mathbf{P}_t^m}$, $\widehat{CE_t^m} = \frac{\widehat{c_t^m}}{(\widehat{\mathbf{P}_t^m})'\mathbf{X}_t}$ is the ratio of optimal cost divided by actual shadow cost, and gauges the efficiency of DMU t in producing output m. Likewise, the associated weight $\widehat{w_t^m} = \frac{(\widehat{\mathbf{P}_t^m})'\mathbf{X}_t}{\widehat{\mathbf{P}_t}'\mathbf{X}_t}$ represents the share of the total shadow cost $\widehat{\mathbf{P}_t}'\mathbf{X}_t$ allocated to output m. Both values $\widehat{CE_t^m}$ and $\widehat{w_t^m}$ are $\in [0,1]$, with $\sum_{m=1}^{M} \widehat{w_t^m} = 1$. As we discuss in our real-world traffic control results, this decomposition substantially leverages the power of the efficiency analysis, particularly in combination with the hourly efficiency measurement.

4.3 Empirical set-up and data

4.3.1 The railway traffic control process

We examine hourly efficiency in railway traffic control, a prime example of multi-output and time-varying 24/7 services. Our empirical model was built in an iterative way, in close collaboration with a panel of Infrabel railway experts. The panel consisted of experts from the operations, human resources, management accounting, and IT departments. When necessary, additional expertise was invoked. In addition, our modelling draws on previous research by Roets and Christiaens (2015), and is in line with the railway traffic control principles as discussed in a world-wide perspective by Pachl (2009) and in a European perspective by Van de Velde et al. (2012).

We define railway traffic control as the combination of signalling activities (i.e., the authorization of train movements), real-time traffic management (i.e., decision making to ensure a fluent and safe traffic flow), and safety actions (e.g. protection of maintenance sites). For a fair efficiency comparison, we only consider control centres equipped with the so-called Automatic Route Setting. This system automatically sets the train route when a train approaches a signal (Pachl, 2009). At Infrabel, Automatic Route Setting is now being gradually rolled out in the traffic control centres. Table 4.1 displays the input and output variables of the empirical model, and table 4.2 presents the descriptive statistics. The inputs consist of the number of staff aligned in the control centre during the hour under evaluation. The local management of the traffic control centre has no control over the exogenously determined outputs but it holds, within the limits of its own authority, responsibility for the optimal alignment of their resources with these outputs.

Table 4.1: Input and output variables

Type	Name	Definition
input	OPER	Number of operators
input	SURV	Number of surveillance staff
output	MOVE	Weighted number of train and local movements (at each signal)
output	ADAPT	Weighted number of non-safety interventions (by standard times)
output	SAFETY	Weighted number of safety interventions (by standard times)

Figure 4.1 visualizes the internal structure of the traffic control production process. The outputs, inputs, and their relation are defined as follows:

- *Output MOVE* considers the number of train movements as well as local movements called '*shunting*' (Pachl, 2009), and is calculated by counting the number of passages at each signal commanded by the control centre. In agreement with the panel of Infrabel railway experts, the movements are weighted by a factor 2 if the signal is opened manually, and a factor of 1 if opened automatically.

- *Output ADAPT* is the time weighted sum of scheduled and unscheduled adaptations to the traffic flow. Examples include merging or splitting trains, re-routing of trains, or special procedures at single-track lines. The weight for each activity is based on their standard execution time in seconds, and was determined by the Infrabel expert panel.

- *Output SAFETY* is the time weighted sum of registered traffic control activities related to safety interventions. Examples of safety procedures are the protection of track maintenance sites through safety locks in the signalling system, or launching safety procedures at level crossings. The weights are analogously defined as for the output ADAPT.

- *Input OPER* is the number of operators, which are responsible for signalling and monitoring activities (captured by the output MOVE) and performing operations related to adapting the traffic flow (ADAPT). Consequently, we model the OPER input as being *shared* by the two outputs MOVE and ADAPT.

- *Input SURV* is the number of surveillance staff, responsible for monitoring and managing the traffic (i.e., taking real-time decisions in case of delays or incidents. In high-density railway networks, numerous traffic control decisions often need to be taken very quickly, i.e. within a few minutes (Van de Velde et al., 2012). Surveillance staff can perform the same actions as the operators. They can instruct operators to perform signalling or adaptation actions, or execute these themselves. The surveillance staff does, however, possess an extended set of skills: they are also authorized to carry out safety procedures (output SAFETY). We therefore model the SURV input as being *shared* by all three outputs.

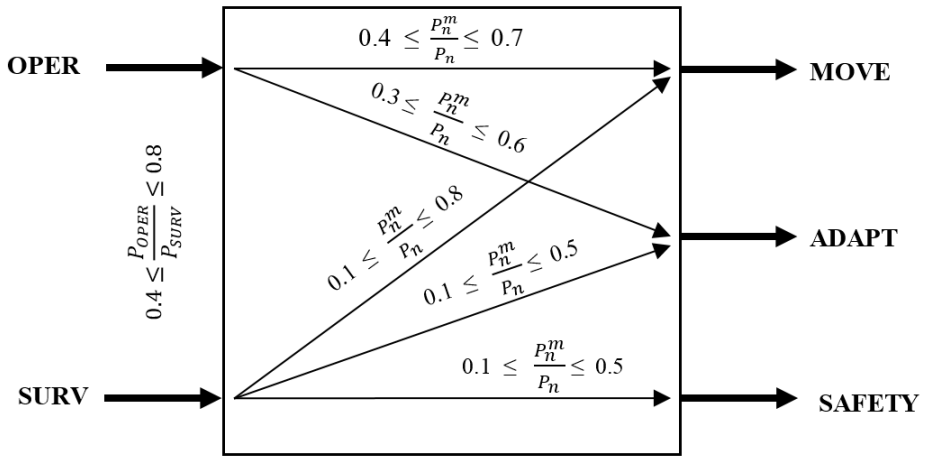


Figure 4.1: Internal structure of the hourly railway traffic control DMU

In case of a major incident, disrupting the entire traffic control zone for an extended period of time (e.g. train derailment or heavy snowfall), the proposed traffic control process is no longer valid, and a separate ad-hoc analysis is warranted. Such cases can easily be identified and omitted from the analysis. The study of major incidents is beyond the scope of this paper.

Applying the cost allocation restricted multi-output approach, we can formally '*look inside the black box*' of the traffic control centres by incorporating production and input allocation information into the model. First, in close consultation with the experts, the production trade-off between the OPER and SURV inputs was estimated as $0.4P_{SURV} \leqslant P_{OPER} \leqslant 0.8P_{SURV}$ (i.e. Assurance Region type I restrictions, Thompson et al., 1986). Second, based on their operational experience, the expert panel defined realistic ranges for the input allocations of the shared inputs. For example, the lower and upper bounds for the fraction $\frac{P_{OPER}^{MOVE}}{P_{OPER}}$ express that between 40 and 70% of operator time is budgeted for signalling and monitoring activities. SURV is shared by all three outputs, with the allocation range for the MOVE output being the widest (10-80% window). The main purpose of the lower bounds for $\frac{P_{SURV}^m}{P_{SURV}}$ was to avoid unrealistic cost allocation schemes (assigning zero output-specific shadow prices). Taken together, the input-output relations and their ranges can be summarized in the following

input-output matrix:

	$\mathbf{L}^{MOVE}$	$\mathbf{U}^{MOVE}$	$\mathbf{L}^{ADAPT}$	$\mathbf{U}^{ADAPT}$	$\mathbf{L}^{SAFETY}$	$\mathbf{U}^{SAFETY}$
$OPER$	0.4	0.7	0.3	0.6	0	0
$SURV$	0.1	0.8	0.1	0.5	0.1	0.5

4.3.2 Human error

Depending on the real-time circumstances in the railway traffic control centre, operators and surveillance staff could be inclined or pressured to apply a modus operandi that induces more human error (such as adapting the pace or quality of the executed tasks). To analyse traffic control errors, we examine the human errors detected by the computerized traffic control system. The observed errors are relatively frequent but non safety-critical[1], and relate to the outputs MOVE (e.g. misordering of track signal commands by the traffic controller) and ADAPT (e.g. lateness in the re-routing of trains, mistyping train or track numbers). The nature of these cognitive errors allows to categorize them as attention failures (Reason, 1990).

Our analysis of human error relates closely to the literature on including undesirable outputs and environmental variables into the efficiency analysis (see, e.g., Cooper et al., 2011, and Daraio and Simar, 2007). However, we opted for keeping human error outside our efficiency analysis, and to analyse the relationship between multi-output efficiency (and its components) and human error in a separate regression analysis.

First, we do not consider human error as an undesirable output. This would imply an implicit trade-off between (overall) efficiency and human error, which may be contrary with managerial objectives (see, e.g., Sherman and Zhu (2006) for a discussion on quality as additional element of performance). In addition, to avoid unbalanced weighting schemes (assigning no or very low importance to a human error output), a priori preference information on attention failures is warranted (Thanassoulis et al., 1995; Shimshak et al., 2009). Further, we only have information on human errors that relate to non-safety errors, making the multi-output efficiency framework incomplete if only the currently observed attention failures are considered. Still, given the potential impact of attention failure on safety, we acknowledge

[1]Data on safety-related human errors (SAFETY output) is not available.

that attention failure could be considered as undesirable output if information is available on the appropriate (social) weights for attention failures and safety-related human errors. Second, we do not include human error as an environmental variable, as human errors are discretionary for the railway traffic control centres. As such, controlling for human error as environmental variable could have perverse influences on the efficiency estimation.

Therefore, to examine and quantify potential relationships between human error and multi-output efficiency, we estimate a series of Probit regression models. We specify the presence of human error as a binary response variable, and (overall or components of) the multi-output efficiency as independent variables. Also, in agreement with the Infrabel experts, we identified the total number of movements and the level of signal automation as two important control variables in predicting error occurrence. The total number of movements is defined as the unweighted number of train and local movements at each signal, while the level of signal automation is calculated as the ratio of automatically opened signals on the total number of signal openings (manual and automatic), multiplied by 100. The variable capturing signal automation is considered as a reasonable proxy for the overall level of traffic control automation, which includes both signal automation and automated route setting. See table 4.2 for descriptive statistics. Overall, we have a 0.38 probability of an occurrence of at least one human error in an hourly interval. On average, there are 155 movements per hour, and 77 percent of signals openings are automated.

4.3.3 24/7/365 data

Aiming to provide actionable insights in staffing levels, work schedules, and human error, we analyse efficiency and error occurrence at a highly disaggregate level. For each single hour of the year 2015, the necessary datasets were generated (allowing for a 24/7/365 analysis). Given the large data volumes, the collection and preparation of the data was performed by means of a Business Intelligence application. The concept behind the custom-developed tool is schematised and discussed in Appendix. The tool unlocked the full potential of the expert panel, especially during the iterative phases of the model building and face-validation: the experts had access to all intermediate versions of the application, and the tool actively supported every model building session (e.g. intuitive assumptions were instantly checked against the available data). In total, 83,607 DMUs have been evaluated for the year 2015.

Table 4.2 presents the descriptive statistics of the dataset.

Table 4.2: Descriptive Statistics (full year 2015)

	Mean	Median	St. Dev.	Minimum	Maximum
OPER	3.03	3.00	1.45	1.00	7.00
SURV	4.62	5.00	1.59	1.00	10.00
MOVE	192.08	153.00	157.09	0.00	1,046.00
ADAPT	805.90	591.00	777.74	0.00	15,037.00
SAFETY	154.93	82.00	222.30	0.00	3,843.00
Human error occurrence	0.38	0	0.48	0	1
Unweighted movements	155.51	128	119.90	0	649
% Auto signalled movements	76.74	78.63	14.31	0	100

4.4 Results and discussion

We structure this section into three subsections. First, we present the general efficiency results and analyse hour-of-day and day-of-week effects. Next, we discuss the pattern exhibited by the binding cost allocation restrictions. Finally, we examine the relationships with human error.

4.4.1 Multi-output efficiency

Table 4.3 presents the average overall efficiency $\widehat{CE}$ and its decomposition in weighted output-specific efficiency scores $\widehat{w^m}.\widehat{CE^m}$. We find on average rather low efficiency levels, with an average overall efficiency of 0.578. As we discuss in the following sections, the low scores can be due to the large variations in hourly efficiency, and the inflexible scheduling of work shifts. According to the Infrabel experts, there could also be an impact of several traffic control centres which are still in expansion (i.e. almost fully dimensioned in technical and human resources, but not yet managing the entire traffic control area). The output-specific efficiency decomposition reveals a high impact of the MOVE efficiency levels, a moderate influence of the ADAPT efficiencies, and low contribution of the SAFETY output to the

overall traffic control efficiency.

Table 4.3: Overall and weighted output-specific efficiency scores

Efficiency	Mean	Median	St. Dev.	Minimum	Maximum
Overall	0.578	0.530	0.178	0.264	1.000
MOVE	0.313	0.286	0.139	0.068	0.781
ADAPT	0.173	0.147	0.103	0.033	0.549
SAFETY	0.092	0.067	0.075	0.014	0.441

The empirical analysis of the hourly efficiency levels allows for a detailed ex post evaluation of staffing levels and schedules. At Infrabel, traffic control staff is aligned 24/7, in a shift schedule with limited flexibility (Van den Bergh et al., 2013): shifts are non-overlapping, with a fixed 8-hour length and fixed starting times at 06:00, 14:00 and 22:00 (i.e., early, late, and night shift). In order to gain insight in the consequences of this limited scheduling flexibility, we examine the hour-of-day and day-of-week variations in traffic control efficiency.

Figure 4.2 presents, for each day of the week and using the 24-hour clock, the hourly profile of the average overall efficiency scores $\widehat{CE}$. Efficiencies are expressed in percentages instead of values between 0 and 1. In order to align visualizations with the current scheduling principles, the graph starts at 06 hours, i.e. the first hour of the early shift. Note that the night shift contains a change of day for the traffic control team involved (e.g. a team starting its shift on Friday 22:00, ends the shift on Saturday 06:00).

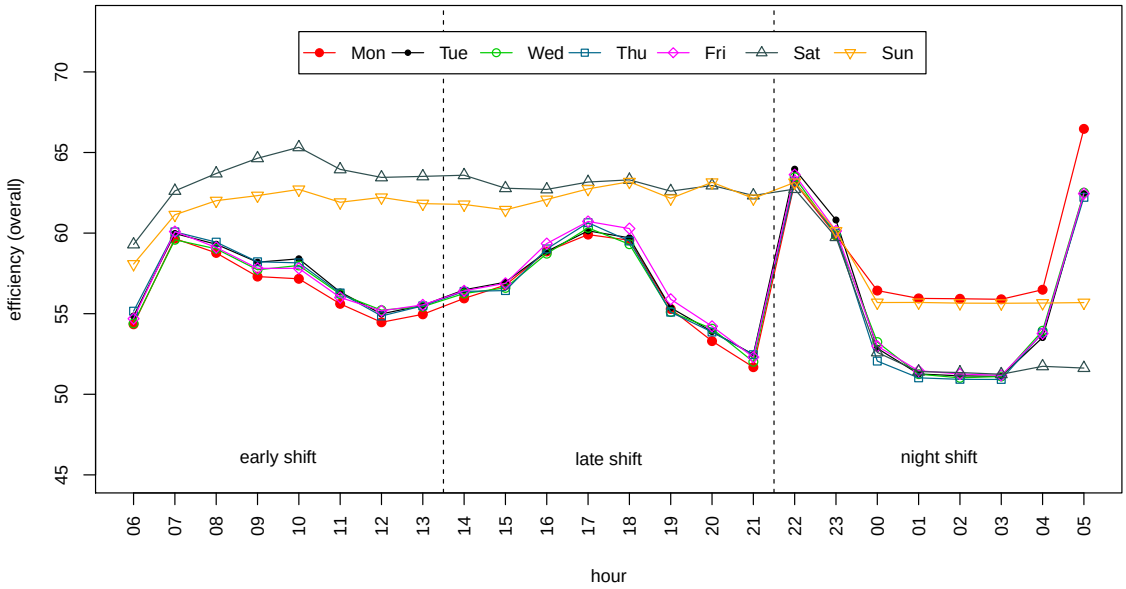


Figure 4.2: Average hourly overall efficiency, for each day of the week

Clearly, the average efficiency levels exhibit considerable variations, with e.g. a 15 percentage point gap between the highest and lowest hourly efficiency. We start by further examining the efficiency levels by looking within the early and late shift. During the working week (Mondays to Fridays), two peaks characterize the efficiency profile. The peaks, situated around 07 and 17 hours, emerge during the morning and evening '*rush hours*' of the railway traffic. After the around 60% efficiency level at 17 hours, traffic control efficiency gradually declines, to end 8 percentage points lower at 21 hours. During the weekend, as can be expected on the basis of the weekend railway timetable, the rush hour effect is absent.

Within the night shift, the efficiency variations paint a more diverse picture. The transition to this shift generates a spike in efficiency for the weekdays (a 12 percentage point jump at 22 hours). Both weekday and weekend efficiencies then gradually decline, and reach a bottom level during the small hours of the night. However, starting from midnight, the efficiency patterns for Sundays and Mondays stabilize at a higher level than the rest of the week (more than 4 percentage points higher). Starting at 04 hours, efficiency improves again, with a considerably lower level on Sundays and Saturdays.

In addition to the hour-of-day and day-of-week effects, our framework also allows to investigate on output-specific performances. By virtue of the applied multi-output methodology, we can further disentangle the efficiency scores and pinpoint the operational reasons underlying the observed efficiency patterns. For the sake of compactness, we focus on the working week. Figure 4.3 visualizes the hourly profiles for each of the weighted output-specific efficiencies $\widehat{w^m}.\widehat{CE^m}$.

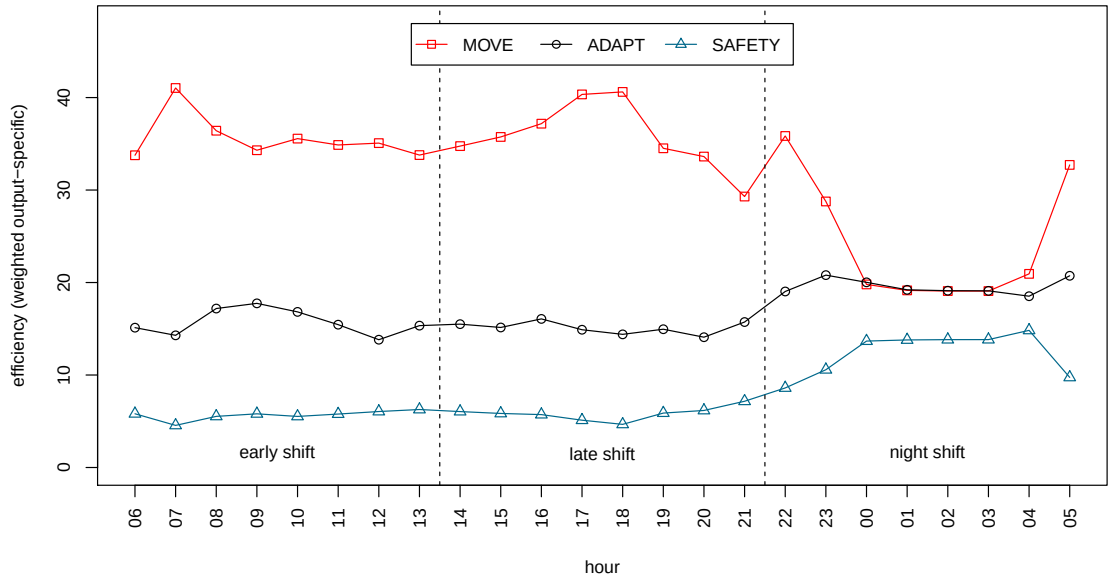


Figure 4.3: Average weighted output-specific efficiency (Monday to Friday)

The hourly efficiency pattern for the MOVE output clearly follows the rhythm of the railway timetable, and explains the first two peaks in the around-the-clock efficiencies. The third overall efficiency peak however, observed at the transition from late to night shift, is a combination of effects from the 3 different output production technologies, and is considerably influenced by the increase of ADAPT (non-safety interventions) and SAFETY (safety interventions, which in practice - and in our model - can only be performed by the surveillance staff SURV). Fully in line with Infrabel expert intuition, a more in-depth analysis reveals that the nightly ADAPT and SAFETY efficiencies are mainly driven by infrastructure maintenance works (which mainly take place during the night). Finally, the peak in overall efficiency at 05 hours is mainly a consequence of the sharp rise in MOVE efficiency.

4.4.2 Binding constraints

Table 4.4 presents the percentage of observations that attain the cost allocation limits defined by the lower and upper bound vectors $\mathbf{L}^m$ and $\mathbf{U}^m$ (i.e., for which the cost allocation restrictions imposed by constraint $C-5$ of the LP model are binding). The highest share of binding constraints is related to the MOVE output, with a share of 44% in the OPER-MOVE relation. In other words, in order to increase overall efficiency, these DMU would have to allocate more than the foreseen maximum of 70% of their OPER resources to the MOVE output. The SURV-MOVE relation exhibits a lesser amount of binding constraints, which might reflect the less restrictive allocation ranges for this input (ranging between 10% and 80%), and the presence of the third output SAFETY. Input cost allocations to SAFETY are never influenced by binding restrictions. This in line with the safety policy at Infrabel, reserving sufficient time for safety procedures.

Table 4.4: Percentage of binding constraints for each input-output relation

	$\mathbf{L}^{MOVE}$	$\mathbf{U}^{MOVE}$	$\mathbf{L}^{ADAPT}$	$\mathbf{U}^{ADAPT}$	$\mathbf{L}^{SAFETY}$	$\mathbf{U}^{SAFETY}$
OPER	9	44	44	9	-	-
SURV	0	20	20	2	37	0

4.4.3 Multi-output efficiency, binding constraints and human error

We analyse the relationship between (the components of) multi-output efficiency and human error by a probit regression. Table 4.5 shows the average marginal effects of our regressions with as dependent variable a dummy that indicates whether there was any registered human error concerning non-safety operations during the hourly interval[2]. We include in all regressions fixed effects for the hour of the day, the day of the week, the month and the railway control centres. Further, we control for the number of movements and the level of automation.

Table 4.5, column 1, shows a clear positive relation between our estimated multi-output efficiency and the probability of erroneously executing tasks. A 0.1 unit increase in multi-output efficiency is associated with a 1.55 percentage points higher probability of human

[2]Observations without any train movements were dropped for the purpose of this regression analysis.

error. Stated differently, lower levels of inputs, given the outputs, are overall associated with more probability for attention failure. In column 2, we focus on the respective output-specific efficiencies $\widehat{CE^m}$. As the reported human errors concern non-safety operations, we focus on the outputs MOVE and ADAPT. We find that it is in particular the efficiency of output ADAPT that correlates positively to human error. A 0.1 unit higher efficiency for the output ADAPT is associated with a 1.92 percentage points higher probability of human error. This makes sense, as it is in the output ADAPT that OPER and SURV are confronted with a highly variable work load[3]. As such, we find that performing highly variable tasks with fewer inputs goes together with more human error. For output MOVE, which concerns more stable operations and tasks that can be more easily planned, we find no significant relation between efficiency and human error. In column 3, we include information on OPER allocation to the output ADAPT. We include the input allocation as a share, and dummy variables for the binding input allocations. We do not find indications for a relation between the OPER allocation share to ADAPT and human error. However, concerning binding constraints and human error, we find a robust and significant association. Allocations of OPER to ADAPT that reach the upper bound – indicating low input levels, given the output – are found to associate with a 6.66 percentage point higher probability of human error. This relation also holds for SURV (column 4), for which we find a 3.32 percentage point higher probability of human error when the allocation of SURV reaches its upper bound. Stated differently, our components of multi-output efficiency provide valuable information on the risk for human error, a proxy for attention failures.

In sum, we find of a clear and robust relation between human error and respectively the efficiency of production tasks that imply highly variable tasks (i.e., output ADAPT) and respectively the binding nature of input allocations. These findings show that the value of multi-output efficiency analysis at a disaggregated level goes beyond the scope of efficiency considerations. As attention failures could impact the safety levels of railway transportation, we provide decision makers a tool to assess multi-output efficiency at an hourly level, while highlighting and quantifying potential risks for safety.

[3]This is reflected in the higher standard deviations for ADAPT, see the descriptive statistics in table 4.2.

Table 4.5: Average marginal effects of probit regression on human error occurrence

	(1)	(2)	(3)	(4)
Multi-output efficiency	0.155**			
	(0.0175)			
Output-specific efficiency (MOVE)		-0.0272	-0.0523	-0.0399
		(0.0181)	(0.0295)	(0.0299)
Output-specific efficiency (ADAPT)		0.192**	0.184**	0.160**
		(0.0154)	(0.0287)	(0.0297)
Allocation of OPER to ADAPT: binding (upper bound)			0.0666**	0.0434**
			(0.0135)	(0.0152)
Allocation of OPER to ADAPT: binding (lower bound)			-0.0130	0.0443
			(0.0236)	(0.0284)
Allocation of OPER to ADAPT			-0.181	0.0241
			(0.115)	(0.137)
Allocation of SURV to ADAPT: binding (upper bound)				0.0332*
				(0.0158)
Allocation of SURV to ADAPT: binding (lower bound)				0.0110
				(0.00694)
Allocation of SURV to ADAPT				0.111**
				(0.0416)
Unweighted movements	0.000992**	0.000977**	0.000973**	0.000971**
	(3.39e-05)	(3.38e-05)	(3.37e-05)	(3.40e-05)
% Auto signalled movements	-0.00445**	-0.00438**	-0.00432**	-0.00432**
	(0.000133)	(0.000132)	(0.000132)	(0.000132)
Hour-of-day fixed effects	Yes	Yes	Yes	Yes
Day-of-week fixed effects	Yes	Yes	Yes	Yes
Month fixed effects	Yes	Yes	Yes	Yes
Railway control centre fixed effects	Yes	Yes	Yes	Yes
pseudo R^2	0.1715	0.1721	0.1721	0.1722
Observations	82,110	82,110	82,110	82,110

Note: Robust standard errors in parentheses.** $p < 0.01$, * $p < 0.05$.

4.5 Conclusion

In this paper, we propose and implement a multi-output cost efficiency framework with cost allocation restrictions, and relate the obtained efficiency scores with human error. As such, we assess both productive efficiency and its relationship with safety. The multi-output approach and the exceptionally disaggregated data allow us to pinpoint staff schedule inefficiencies, while highlighting potential safety risks. Using railway traffic control as an unique

testing ground, the concept is developed from start to finish in close cooperation with experts from Belgian railways. As such, the tool directly responds to the increasing pressure on Europe's railway infrastructure managers to uplift their efficiency, without sacrificing safety levels.

The advocated multi-output cost efficiency framework with proportional cost allocation restrictions has the following features. It allows for a transparent and flexible incorporation of a priori process information in the model, which considerably strengthens the credibility and acceptance of the results. Unique is that the cost allocation restrictions define all input types, including output-specific, shared, joint and sub-joint inputs. The proportional allocation bounds allow for a convenient intuitive interpretation, which strongly simplifies the communication and interaction with business experts. As such, by carefully adapting our methodology to address the traffic control problem at hand, we have conceived a more general framework, able to handle a wide range of real-life applications. Providing quantitative insights in hour-of-day, day-of-week, and output-specific efficiency variations, the framework empowers management to focus their attention to the most prominent staffing efficiency issues, and optimize their staffing levels and work shift patterns on an ex post basis. The approach can readily be extended with a straightforward '*what-if*' analysis, by using planned or mathematically optimized staff schedules as an input for the efficiency calculations.

Our empirical efficiency results reveal a 15 percentage point gap between the (average) highest and lowest hourly efficiency levels. This suggests that management should evolve from a relatively inflexible and '*one-size-fits-all*' scheduling philosophy, consisting of non-overlapping fixed 8-hour shifts, to a more customized approach (e.g. by gradually changing team size and composition, revise shift length and starting times, and scheduling overlapping shifts). Overall, we find no evidence against the idea that sufficient time is allocated to perform safety operations.

Our contribution to the literature extends above and beyond the scope of traditional efficiency methods, by uncovering a significant relationship between human error and the multi-output efficiencies. We observe a clear positive relationship between human error and the efficiency of production tasks of a more variable and unpredictable nature. For these production tasks, reaching the upper boundaries of the modelled input-output cost alloca-

tion restrictions also significantly impacts error probability. All revealed relationships are quantified through probit marginal effect estimations. The finding of a significant relationship between efficiency components and potentially more hazardous outcomes gives rise to policy recommendations stretching beyond the sheer efficiency perspective. More specifically, when examining the possible detrimental effects of efficiency changes on safety levels, decision makers should focus on the staff allocations towards outputs with a highly variable and unpredictable workload. The circumstances during the real-time execution of their (multi-output) activities could motivate staff to choose for a modus operandi which leads to impaired attention levels, which in turn could impact both non-safety and safety-related outputs.

For safety reasons, and fully in line with the above, Belgian railways are currently establishing a new work organisation, in which the responsibilities for the safety actions will be assigned to dedicated '*safety controllers*', while all non-safety tasks will be handled by '*traffic controllers*'. The developed framework can then be applied to assess the impact of the reorganisation on efficiency, and examine the empirical link with human error. A limitation of our railway traffic control application however, is the absence of error measures for the safety actions. Therefore, efforts are currently underway to define measures on safety-related human errors, and extract the corresponding data out of the traffic control systems. As such, the real-life implementation of the framework into Belgian railway traffic control will continue to allow for extensive practitioner-based validation and feedback. Finally, given the large volumes of data involved and the support of a concomitant Business Intelligence tool, our research will also foster further research on developing DEA tools for large-scale production data.

Acknowledgements

We are grateful to Laurens Cherchye and Bram De Rock for useful comments. The authors would like to thank Infrabel for funding this research, and providing the necessary railway data and expertise. Special thanks goes to the management and experts of the Traffic Operations department, the Methods department, the Financial Controlling department, as well as to Performance Data division and the Business Analytics team. It is to be noted that the views expressed in this paper are those of the authors and do not necessarily reflect the

opinions of Infrabel.

Appendix: Business Intelligence Tool

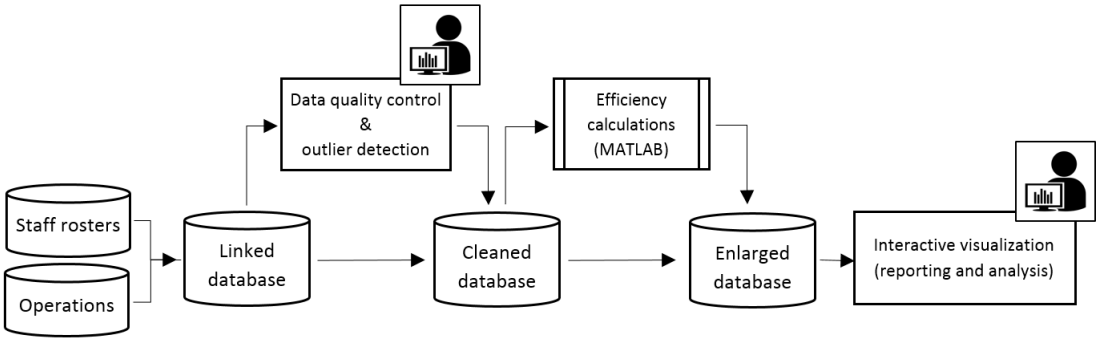


Figure 4.4: Application of the Business Intelligence tool

The custom-developed Business Intelligence tool is code-named 'CRIPTON': Comprehensive Railway Infrastructure Performance Tools for Operations on the Network. The applied server-based Business Intelligence software is QlikView. The application loads the available micro-level data from the staff rostering and traffic control databases, and links the two sources of information into a single database. The staff roster data contains the actual and not the scheduled working time. The tool calculates the datasets for the multi-output efficiency model (and the human error analysis) by aggregating the data at 24/7/365 level. For each traffic control centre, and for each hour and each day of the year, a DMU is generated. For example, the DMU '*Brussels-2015-10-27-10h*' covers the inputs and outputs for the Brussels traffic control centre, on October 27, from 10:00 until but not including 11:00 (24-hour clock). As the number of traffic control centres equipped with Automatic Route Setting gradually extends, our sample considered 8 traffic control centres in January 2015 (i.e. 31x24x8 = 5,952 DMUs) and 11 traffic control centres in December (i.e. 31x24x11 = 8,184 DMUs). All observations are pooled into a single full year dataset.

Using customized graphs and tables, the data quality was systematically checked for the entire database as well as for the generated DMU dataset. This allows for an in-depth data verification and a visual detection of outliers in the DMU data. The outcome of the data quality control process was discussed with the expert panel during preparatory meetings, after which the agreed data cleaning was performed. Several railway strike days were eliminated from the data set, as well as 43 observations considered as outliers (e.g. major incidents)

or with errors in the data. The full year DMU dataset is then exported from the cleaned database to a MATLAB environment for the efficiency calculations. The probit regressions are performed with STATA. After importation from MATLAB the Business Intelligence application adds the efficiency results to the cleaned database, and links each DMU with the corresponding operational micro-level data.

The Business Intelligence application has demonstrated its merits during the entire research process. The initial collection of the data, the creation of a linked database, the verification of data quality, the specification of the production model and finally the face-validation of the results have all been supported or enabled by the application.

References

Allen, R., Athanassopoulos, A., Dyson, R. G., and Thanassoulis, E. (1997). Weights restrictions and value judgements in data envelopment analysis: evolution, development and future directions. *Annals of Operations Research*, 73:13–34.

Beasley, J. E. (1995). Determining teaching and research efficiencies. *Journal of the operational research society*, pages 441–452.

Castelli, L., Pesenti, R., and Ukovich, W. (2010). A classification of DEA models when the internal structure of the decision making units is considered. *Annals of Operations Research*, 173(1):207–235.

Charnes, A., Cooper, W. W., and Rhodes, E. (1978). Measuring the efficiency of decision-making units. *European journal of operational research*, 2(6):429–444.

Cherchye, L., De Rock, B., and Hennebel, V. (2017). Coordination efficiency in multi-output settings: a DEA approach. *Annals of Operations Research*, 250(1):205–233.

Cherchye, L., De Rock, B., and Walheer, B. (2015). Multi-output efficiency with good and bad outputs. *European Journal of Operational Research*, 240(3):872–881.

Cherchye, L., Demuynck, T., Rock, B., and Witte, K. (2014). Non-parametric analysis of multi-output production with joint inputs. *The Economic Journal*, 124(577):735–775.

Cherchye, L., Rock, B. D., Dierynck, B., Roodhooft, F., and Sabbe, J. (2013). Opening the "black box" of efficiency measurement: input allocation in multioutput settings. *Operations Research*, 61(5):1148–1165.

Cook, W. D., Chai, D., Doyle, J., and Green, R. (1998). Hierarchies and groups in DEA. *Journal of Productivity Analysis*, 10(2):177–198.

Cook, W. D. and Hababou, M. (2001). Sales performance measurement in bank branches. *Omega*, 29(4):299–307.

Cook, W. D., Hababou, M., and Tuenter, H. J. (2000). Multicomponent efficiency measurement and shared inputs in data envelopment analysis: an application to sales and service performance in bank branches. *Journal of Productivity Analysis*, 14(3):209–224.

Cook, W. D. and Kress, M. (1999). Characterizing an equitable allocation of shared costs: A DEA approach. *European Journal of Operational Research*, 119(3):652–661.

Cook, W. D. and Seiford, L. M. (2009). Data envelopment analysis (DEA)–thirty years on. *European Journal of Operational Research*, 192(1):1–17.

Cook, W. D. and Zhu, J. (2011). Output-specific input-assurance regions in DEA. *Journal of the Operational Research Society*, 62(10):1881–1887.

Cook, W. D. and Zhu, J. (2014). *Data Envelopment Analysis: A handbook of modeling internal structure and network*, volume 208. Springer.

Cooper, W. W., Seiford, L. M., and Zhu, J. (2011). *Data Envelopment Analysis: Handbook on Data Envelopment Analysis*. Springer.

Daraio, C. and Simar, L. (2007). *Advanced robust and nonparametric methods in efficiency analysis: Methodology and applications*. Springer Science & Business Media.

Despić, O., Despić, M., and Paradi, J. C. (2007). DEA-r: ratio-based comparative efficiency model, its mathematical relation to DEA and its use in applications. *Journal of Productivity Analysis*, 28(1-2):33–44.

European Directive 2012/34/EU (2012). Directive 2012/34/EU of the European Parliament and of the Council of 21 november 2012 establishing a single european railway area. *Official Journal of the European Union, 55*.

European Directive 2016/798/EU (2016). Directive 2016/798/EU of the European Parliament and of the Council of 11 may 2016 on railway safety. *Official Journal of the European Union, 59*.

Färe, R. and Grosskopf, S. (1996). Productivity and intermediate products: A frontier approach. *Economics letters*, 50(1):65–70.

Färe, R. and Grosskopf, S. (2000). Network DEA. *Socio-economic planning sciences*, 34(1):35–49.

Fare, R., Grosskopf, S., and Whittaker, G. (2007). Network DEA: Chapter 12 in modelling data irregularities and structural complexities in data envelopment analysis. j. zhu and wd cook.

Fried, H. O., Lovell, C. K., and Schmidt, S. S. (2008). *The measurement of productive efficiency and productivity growth*. Oxford University Press.

Halme, M. and Korhonen, P. (2000). Restricting weights in value efficiency analysis. *European Journal of Operational Research*, 126(1):175–188.

Hoopes, B. J. and Triantis, K. P. (2001). Efficiency performance, control charts, and process improvement: complementary measurement and evaluation. *Engineering Management, IEEE Transactions on*, 48(2):239–253.

Jain, S., Triantis, K. P., and Liu, S. (2011). Manufacturing performance measurement and target setting: A data envelopment analysis approach. *European Journal of Operational Research*, 214(3):616–626.

Joro, T. and Korhonen, P. J. (2015). *Extension of Data Envelopment Analysis with Preference Information*, volume 218. Springer.

Leibenstein, H. (1966). Allocative efficiency vs." x-efficiency". *The American Economic Review*, pages 392–415.

Leibenstein, H. (1973). Competition and x-efficiency: Reply. *Journal of Political Economy*, 81(3):765–777.

Lozano, S. (2015). A joint-inputs network DEA approach to production and pollution-generating technologies. *Expert Systems with Applications*, 42(21):7960–7968.

Oliveira, M., Camanho, A., and Gaspar, M. (2014). Enhancing the performance of quota managed fisheries using seasonality information: the case of the portuguese artisanal dredge fleet. *Marine Policy*, 45:114–120.

Pachl, J. (2009). *Railway operation and control*. VTD Rail Publishing, Mountlake Terrace (USA).

Podinovski, V. (2004). Production trade-offs and weight restrictions in data envelopment analysis. *Journal of the Operational Research Society*, pages 1311–1322.

Podinovski, V. V. (2016). Optimal weights in DEA models with weight restrictions. *European Journal of Operational Research*, 254(3):916–924.

Podinovski, V. V. and Bouzdine-Chameeva, T. (2013). Weight restrictions and free production in data envelopment analysis. *Operations Research*, 61(2):426–437.

Podinovski, V. V. and Bouzdine-Chameeva, T. (2015). Consistent weight restrictions in data envelopment analysis. *European Journal of Operational Research*, 244(1):201–209.

Reason, J. (1990). *Human error*. Cambridge university press.

Roets, B. and Christiaens, J. (2015). Evaluation of railway traffic control efficiency and its determinants. *European Journal of Transport and Infrastructure Research*, 15(4):396–418.

Salerian, J. and Chan, C. (2005). Restricting multiple-output multiple-input DEA models by disaggregating the output–input vector. *Journal of Productivity Analysis*, 24(1):5–29.

Sarrico, C. S. and Dyson, R. (2004). Restricting virtual weights in data envelopment analysis. *European Journal of Operational Research*, 159(1):17–34.

Sherman, H. D. and Zhu, J. (2006). Benchmarking with quality-adjusted DEA (q-DEA) to seek lower-cost high-quality service: evidence from a us bank application. *Annals of Operations Research*, 145(1):301–319.

Shimshak, D. G., Lenard, M. L., and Klimberg, R. K. (2009). Incorporating quality into data envelopment analysis of nursing home performance: a case study. *Omega*, 37(3):672–685.

Thanassoulis, E., Boussofiane, A., and Dyson, R. (1995). Exploring output quality targets in the provision of perinatal care in england using data envelopment analysis. *European Journal of Operational Research*, 80(3):588–607.

Thompson, R. G., Langemeier, L. N., Lee, C.-T., Lee, E., and Thrall, R. M. (1990). The role of multiplier bounds in efficiency analysis with application to kansas farming. *Journal of Econometrics*, 46(1):93–108.

Thompson, R. G., Singleton Jr, F., Thrall, R. M., and Smith, B. A. (1986). Comparative site evaluations for locating a high-energy physics lab in texas. *Interfaces*, 16(6):35–49.

Van de Velde, D., Nash, C., Smith, A., Mizutani, F., Uranishi, S., Lijesen, M., and Zschoche, F. (2012). Eves-rail—economic effects of vertical separation in the railway sector. *Report for the Community of European Railway and Infrastructure Companies*.

Van den Bergh, J., Beliën, J., De Bruecker, P., Demeulemeester, E., and De Boeck, L. (2013). Personnel scheduling: A literature review. *European Journal of Operational Research*, 226(3):367–385.

Vázquez-Rowe, I. and Tyedmers, P. (2013). Identifying the importance of the "skipper effect" within sources of measured inefficiency in fisheries through data envelopment analysis (DEA). *Marine Policy*, 38:387–396.

Wong, Y.-H. and Beasley, J. (1990). Restricting weight flexibility in data envelopment analysis. *Journal of the Operational Research Society*, pages 829–835.

Chapter 5

GENERAL CONCLUSIONS AND FUTURE RESEARCH

5.1 Managerial challenges and opportunities

European railways are **under increasing pressure to raise efficiency without sacrificing safety**. Since 1991, European directives gradually unbundled the railway system into national 'infrastructure managers', and several competing railway undertakings. With the European Directive 2012/34/EU (2012), the European Union is pursuing the development of a competitive Single European Railway Area. It considers railway infrastructure as a natural monopoly, and as such urges the infrastructure managers to reduce costs. At the same time, railway safety levels should be maintained and even improved where practicable (European Directive 2016/798/EU, 2016).

In support of these challenges, this dissertation focuses on railway traffic control, a core railway infrastructure activity which leans heavily on efficiency and safety to improve performance. European railway traffic control is characterized by a **large-scale technological migration towards computerized control centres** (Wilson and Norris, 2005), which increasingly opens opportunities for in-depth and data-driven research. In the UK, the Office of Rail Regulation has provided a good overview of this long-term technological shift in its international benchmarking study (Civity management consultants, 2013). The report concludes that optimal migration strategies should not only consist of modernizing traffic control centres but also, and in parallel, optimizing staffing levels. More specifically, **railway infrastructure managers should take full advantage of the possibilities offered by the**

centralized and therefore larger traffic control centres, by crafting an appropriate Human Resource strategy. This includes measures such as more detailed staffing methods and an increased scheduling flexibility.

The performance analysis tools, empirical findings, and policy recommendations, bundled in this dissertation, can support infrastructure managers in tackling these real-world issues. **Rooted in the two distinct disciplines of efficiency estimation and fatigue modelling**, it provides a broader, non-unidimensional answer to the overarching question which has driven this research: how to improve staffing efficiency, while accounting for human factors and safety concerns. As such, it introduces the new research field of railway traffic control efficiency, and deepens insight in the underresearched area of railway traffic controller fatigue. The multidisciplinary nature of the dissertation also allows for a triangulative approach to examine human error, by linking its probability to both staffing efficiency and fatigue risk. The developed models and the empirical findings are supported by a purpose-built Business Intelligence environment, fuelled by real-world Belgian railway data. This actively bridges the gap between researchers and railway experts, and as such substantially leverages the face-validity of the research.

Regardless of the perspective taken (efficiency or fatigue), the presented performance assessments are **based on an ex-post approach**, in which past performance is analysed and processed into useful managerial information. This is in line with the inherent nature of the DEA methodology, which is generally applied as a mathematical programming tool for ex-post efficiency evaluations (see e.g., Banker et al., 1984; Joro and Korhonen, 2015). In contrast, biomathematical fatigue models are usually applied in a predictive sense. The regression-based approach developed in this dissertation follows, however, the recommendation of Dawson et al. (2017) to conceptualize the output of a fatigue risk model as a continuous point-estimate, which needs to be appraised while taking account of workplace specifics. As such, we apply retrospective statistical analysis to simultaneously link observed human errors with the Risk Index and other risk influencing factors. This 'post-implementation surveillance mode' has, in contrast with the normal prospective risk assessment application of fatigue risk models, received very little attention (see ibidem).

5.2 Main policy recommendations

In addition to each paper's individual contributions to the literature, as summarized in chapter 1 and further detailed in the subsequent chapters of this dissertation, there are a number of policy recommendations which merit to be highlighted.

- First, as suggested in the chapter on non-computerised traffic control centres, **it pays to reduce infrastructure complexity**. An asset management strategy, aiming for 'lean infrastructure' not only lowers asset maintenance cost, but also has positive effects on traffic control efficiency.

- Second, railway infrastructure managers **should take full advantage** of the possibilities offered by the **centralized and larger traffic control centres**, by developing **a matching Human Resource strategy**. In parallel with the centralization strategies, this can increase efficiency levels and help to tackle the challenges of the imposed austerity measures and a bothersome 'age pyramid'.

- Third, the 24/7 alignment of staff in the control centres calls for particular **management attention for employee well-being** (e.g. work-life balance) **and safety concerns**. With a specific focus on safety, and as elaborated in the chapter on fatigue risk, the real-world and railway traffic control validation of the Risk Index demonstrates its **ability to evaluate staff schedules on potential fatigue risk issues**. However, results also suggest that safe work schedule design should not exclusively rely on fatigue risk estimations, but also take into account day-of-the-week effects.

- Fourth, by applying the hourly and multi-output performance model, developed in the chapter on the efficiency of computerised traffic control centres, railway infrastructure managers can easily **reveal inefficiencies in their staff schedules**. By adjusting staffing levels and work shift patterns accordingly (an exercise leveraged by the larger team sizes aligned in the computerised centres) efficiency levels can be iteratively improved. When examining the **possibly detrimental effects of efficiency changes on safety levels**, special attention should be given to **workload of a less predictable nature**.

- Fifth, and expanding on the previous two policy recommendations: when measuring

and judging traffic control performances, railway infrastructure management **should be aware of day-of-week and time-of-day effects**. These effects can manifest themselves in both the efficiency and fatigue risk dimensions of performance.

5.3 The 'performance assessment system of the future': some reflections

The ongoing digitization of traffic control systems in Europe has presented an excellent and timely opportunity for the data-driven research presented in this dissertation. Although the approach was pioneered for railway traffic control centres, it can be generalized to similar control centre types, operating in a safety-critical environment and on a 24/7 basis. **Efficient and safe control centre operation is paramount** in a large number of industries: acting as the nerve centre for real-time monitoring and intervention, control centres manage and coordinate air traffic, road traffic, gas pipelines, nuclear power plants, chemical production sites, and many other safety-critical environments.

As showcased in this dissertation, a **comprehensive control centre performance assessment system** should be based on **three pillars**. First, performance should be **approached from different and possibly counterbalancing perspectives**. Second, a **link between workforce and operational data** should be established, and this at different disaggregation levels (both in terms of data as well as production process description). Third, this data link should be **systematic and permanent**, in order to allow sustained performance monitoring and evaluation.

Cogitating and ruminating on a further extension of this concept – to be developed above and beyond the practical constraints of a doctoral research project – a fourth pillar can be added. To be ultimately comprehensive, the current **ex-post** performance measurement should be **complemented with ex-ante forecasts** and **ex-nunc (real-time) performance evaluations**[1] (see figure 5.1).

[1]The terminology being borrowed from performance management approaches in New Public Management, see Van Dooren et al. (2015).

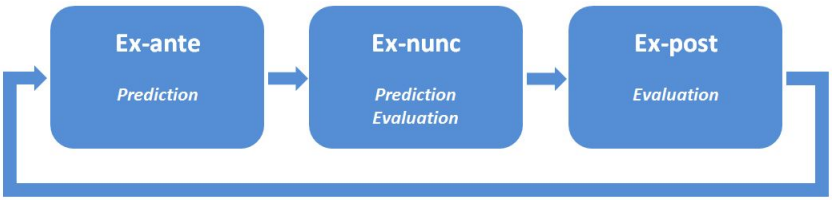


Figure 5.1: Control centre 'performance assessment system of the future'

Evidently, the new time perspectives will lead to new challenges to be addressed, such as the development and validation of predictive performance models, the gathering of real-time data (with, for example, eye-tracking devices or computerised fitness-for-duty tests for fatigue detection purposes), or real-time efficiency estimation issues (the required DEA calculations being oxymoronic). In addition, there is a need to track and explore the potential of emerging real-time (and data-driven) fatigue risk assessment technologies. As indicated by Gander et al. (2011) and Dawson et al. (2012), comprehensive fatigue risk management not only considers an ex-ante 'fatigue reduction' strategy (e.g. through work schedule screening), but also includes 'fatigue proofing' approaches to identify the fatigue-impaired individuals already present on the work floor, and mitigate the likelihood of their fatigue-related errors.

Given the recent technological advances, the (fatigue risk aspects of the) real-time performance component should be further divided into two sub-components: **real-time evaluation** and **real-time prediction**. A promising and implementable real-time (fatigue and stress) *evaluation* technology is behavioural biometrics, based on a machine-learning algorithm for analysing computer keystroke and mouse movement dynamics, while taking account of circadian rhythmicity and task complexity (see Carneiro et al., 2017a,b). Real-time *prediction* of operator fatigue – estimating operator drowsiness for the next upcoming seconds – could rely on the appropriate real-time processing of a level-of-drowsiness (LoD) signals produced by, for example, images of the eye and related ocular parameters. An intriguing approach for doing this is suggested by Ebrahimbabaie and Verly (2017), who have developed signal processing algorithms for modeling LoD signals through a geometric Brownian motion (GBM) random process model and using this model for predicting the future values of the LoD, as well as for predicting other events such as exceeding a predetermined LoD level. Finally, at a conceptual level, the linkages between the three distinct time perspectives of the performance

assessment cycle (i.e., ex-ante, ex-nunc, ex-post), and the relative importance of each of the considered performance dimensions (e.g. real-time fatigue or stress detection being more vital than real-time efficiency estimation) will have to be determined.

5.4 Looking forward: objectives and approaches for follow-up research

As this doctoral research gained momentum and depth, a wide range of opportunities for further research gradually emerged, all with specific theoretical and empirical contributions. As such, this dissertation has laid the groundwork for data-driven follow-up research, geared towards the 'performance assessment system of the future' described above. The stage now gets set for an extensive research program which will not only build on the developed models and insights, but also continues the intensive researcher/railway expert cooperation. **Three distinct but complementary lines of thought** have been developed at this time.

First, fully embracing the different performance dimensions discussed in this dissertation, an **all-encompassing concept** for railway traffic performance will be developed. The objective is to conceptualise a comprehensive performance assessment system, simultaneously capturing and explaining staffing efficiency, fatigue risk, and human error. This allows to evolve from the current 'multidisciplinary' nature of the research to a more 'interdisciplinary' approach. In order to fully take account of the interactions between these different performance dimensions (and possible others to be discerned at a later stage), a 'systems oriented' approach is warranted. Applying **systems thinking as an underlying and pivotal notion**, the follow-up research will be positioned at the unique intersection of the literatures on system performance, efficiency measurement, and fatigue management. As indicated above, the socio-technical framework should be generalizable to a wide range of control centres, operating in a safety-critical environment and on a 24/7 basis.

Second, by diving deeper into the details of the production process, and by **constructing datasets and models at workstation level** instead of control centre level, a direct link between predicted traffic controller fatigue and the performed operations can be established. The inability of work schedule-based fatigue models to **take account of individual char-**

acteristics is however recognized in the literature (e.g. Dawson et al., 2011). One possible solution is to empirically test the recent extension of the Folkard et al. (2007) Risk Index (Fischer et al., 2017, currently under review), which allows for a more personalised risk estimation by considering the operators' circadian chronotype (morningness-eveningness tendency) as a fatigue risk model input. In addition, the analysis could be further strengthened by **lifting the temporal resolution of the Risk Index to the hourly level** (instead of the current work shift average). This would allow to combine the fatigue risk estimations (i.e., chapter 3) with the power of the hourly efficiency calculations (i.e., chapter 4), and this at workstation level (i.e., at a new data and process disaggregation level). Also, by advancing the traffic control performance analysis to the workstation level, the scope of application can be further widened to organisations with only a few (or just one) control centre(s). However, as this level of data disaggregation draws the research nearer to an assessment of individual performance and behaviour, a general caveat is in order. With a frictionless implementation of the performance measurement system in mind, the research will need to continue its non-intrusive approach, and ensure the personal privacy of the individual traffic controllers. Particularly in highly unionized environments – such as the railways or air traffic control – this is critical in terms of user acceptance and satisfaction. Therefore, to mitigate potential implementation issues, an extension of the research efforts towards the literature on 'Electronic Performance Monitoring' systems - and their effects on organizational and individual performance - may be warranted (see, e.g., Alge and Hansen, 2014).

Third, in order to forecast future performance, the above-mentioned ex-post approaches will be **complemented by a predictive (ex-ante) model**. Making full use of the systems thinking paradigm, the complex and dynamic character of the traffic control workstation processes (including the interplay between human, operational, and technological components) will be captured in a **system dynamics simulation** application. The ex-ante predictions generated by the simulations will empower decision makers with a quantitative tool, providing insights in stability, equilibrium states and transition path dynamics. This can reinforce and support decision making, by predicting the strategic and operational outcomes of managerial interventions (e.g. changes in both efficiency and safety levels). In addition, the analysis and visualization of the complex and dynamic processes will lead to improved mental models of the real-world traffic control setting: mental models (and human decision making) are

constrained by cognitive limitations, generally ignore feedback processes and are dynamically deficient (Sterman, 2000). Aiming for an **explicit link of the system dynamics concept with productive efficiency measurement** (Vaneman and Triantis, 2003), the predictive model will be based on the Dynamic Productive Efficiency Measurement framework (Vaneman and Triantis, 2007). Where previous applications of the Dynamic Productive Efficiency Measurement model were oriented towards physical systems, the current model will expand this methodology towards socio-technical systems with an explicit safety-critical component. One of the particular challenges associated with this concept is to develop representations of the disaggregated production structure, which capture and quantify the dynamic behaviour of efficiency, fatigue, and safety.

To summarise, the described **follow-up research** will continue to explore the 'boundaries of acceptable performance' (Rasmussen, 1997), with an initial focus on **ex-post and ex-ante performance assessment**. As in many other safety-critical and complex systems, it is not to be excluded that the increasing pressure for efficiency will also push railway traffic control operations 'closer to the workload and safety boundaries' (Dekker, 2016). Special caution is warranted for the looming dangers of 'decrementalism' (ibid.), where small and gradually implemented productivity gains slowly and almost unnoticeably carve out workload and safety margins. The development and implementation of a comprehensive performance assessment system can help in safeguarding traffic control operations (or any other safety-critical setting) against these progressive performance decrements. As such, the basic research question of this dissertation, i.e. 'how to improve staffing efficiency, while accounting for human factors and safety concerns', continues to drive the research.

References

Alge, B. J. and Hansen, S. D. (2014). Workplace monitoring and surveillance research since "1984": A review and agenda. *The psychology of workplace technology, New York, NY*, pages 209–237.

Banker, R. D., Charnes, A., and Cooper, W. W. (1984). Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management science*, 30(9):1078–1092.

Carneiro, D., Novais, P., Augusto, J. C., and Payne, N. (2017a). New methods for stress assessment and monitoring at the workplace. *IEEE Transactions on Affective Computing*.

Carneiro, D., Pimenta, A., Neves, J., and Novais, P. (2017b). A multi-modal architecture for non-intrusive analysis of performance in the workplace. *Neurocomputing*, 231:41–46.

Civity management consultants (2013). International benchmarking of network rail's operations and support functions expenditure. *Department for Transport and Office of Rail Regulation; London*.

Dawson, D., Chapman, J., and Thomas, M. J. (2012). Fatigue-proofing: a new approach to reducing fatigue-related risk using the principles of error management. *Sleep medicine reviews*, 16(2):167–175.

Dawson, D., Darwent, D., and Roach, G. D. (2017). How should a bio-mathematical model be used within a fatigue risk management system to determine whether or not a working time arrangement is safe? *Accident Analysis & Prevention*, 99:469–473.

Dawson, D., Noy, Y. I., Härmä, M., Åkerstedt, T., and Belenky, G. (2011). Modelling fatigue and the use of fatigue models in work settings. *Accident Analysis & Prevention*, 43(2):549–564.

Dekker, S. (2016). *Drift into failure: From hunting broken components to understanding complex systems*. CRC Press.

Ebrahimbabaie, V. P. and Verly, J. (2017). Discovery that a person's level of drowsiness appears to evolve in time according to a geometric brownian motion (gbm) random process model.

European Directive 2012/34/EU (2012). Directive 2012/34/EU of the European Parliament and of the Council of 21 november 2012 establishing a single european railway area. *Official Journal of the European Union, 55*.

European Directive 2016/798/EU (2016). Directive 2016/798/EU of the European Parliament and of the Council of 11 may 2016 on railway safety. *Official Journal of the European Union, 59*.

Fischer, D., Lombardi, D. A., and Folkard, S. (2017). Extension of risk index. *Under revision*.

Folkard, S., Robertson, K. A., and Spencer, M. B. (2007). A fatigue/risk index to assess work schedules. *Somnologie-Schlafforschung und Schlafmedizin*, 11(3):177–185.

Gander, P., Hartley, L., Powell, D., Cabon, P., Hitchcock, E., Mills, A., and Popkin, S. (2011). Fatigue risk management: Organizational factors at the regulatory and industry/company level. *Accident Analysis & Prevention*, 43(2):573–590.

Joro, T. and Korhonen, P. J. (2015). *Extension of Data Envelopment Analysis with Preference Information*, volume 218. Springer.

Rasmussen, J. (1997). Risk management in a dynamic society: a modelling problem. *Safety science*, 27(2):183–213.

Sterman, J. D. (2000). *Business dynamics: systems thinking and modeling for a complex world*. McGraw-Hill, Boston.

Van Dooren, W., Bouckaert, G., Halligan, J., et al. (2015). *Performance management in the public sector*. Routledge.

Vaneman, W. K. and Triantis, K. (2003). The dynamic production axioms and system dynamics behaviors: the foundation for future integration. *Journal of Productivity Analysis*, 19(1):93–113.

Vaneman, W. K. and Triantis, K. (2007). Evaluating the productive efficiency of dynamical systems. *IEEE Transactions on Engineering Management*, 54(3):600–612.

Wilson, J. R. and Norris, B. J. (2005). Rail human factors: Past, present and future. *Applied ergonomics*, 36(6):649–660.