EJTIR

Issue 15(4), 2015
pp. 396-418
ISSN: 1567-7141
tlo.tbm.tudelft.nl/ejtir

Evaluation of railway traffic control efficiency and its determinants

Bart Roets[1]
Faculty of Economics and Business Administration, Ghent University, Belgium.
Traffic Management & Services department, Infrabel, Belgium.

Johan Christiaens[2]
Faculty of Economics and Business Administration, Ghent University, Belgium.

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 weekend when measuring and analysing traffic control performance.

***Keywords**: bootstrap, data envelopment analysis, efficiency, railway infrastructure, traffic control, two-stage approach.*

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[3] on the establishment of a Single European Railway Area (2012) 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

[1] A: Frankrijkstraat 85, 1060 Brussel, Belgium T: +32 2 525 90 10 F: +32 2 526 38 08 E: bart.roets@ugent.be
[2] A: Sint-Pietersplein 7, 9000 Gent, Belgium T: +32 9 264 35 40 F: +32 9 264 35 92 E: johan.christiaens@ugent.be

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

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 Rails[4] 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, A., Wheat, and Smith G., 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[5], 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 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 (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[6] 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 1800 persons, spread over up to 120 signal boxes (average number in 2013)[7].

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

[4] The British railway infrastructure manager.

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

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

[7] Source: Infrabel data.

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 sub-regions. 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. 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; Mats 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, Wiggenraad, and Wolff (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 UIC InfraCost study (Union Internationale des Chemins de fer, 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[8] within the context of the McNulty Value For Money report (2011), and in a further extension of this study in a benchmarking report on operations and support functions (UK Office of Rail Regulation, 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; NERA, 2006; EUROCONTROL Performance Review Unit, 2011). Recently, two academic articles have been published (Button and Neiva, 2013 and 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 between air and rail traffic control, we refer to Pellegrini and Rodriguez (2013).

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, Seiford, and Zhu, 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

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

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, de la Garza, and Triantis, 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.

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[9]. 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.

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, p. 23). 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

[9] 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 InfraCost (2002), the McNulty Value for Money study (2011), and the UK Office of Rail Regulation benchmarking report (2013).

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[10]. 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 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 the UIC InfraCost study, 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

[10] 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).

largely influenced by macro-economic factors or public service requirements (see e.g. Merkert, Smith, and Nash, 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 (UIC InfraCost study, 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 (UK Office of Rail Regulation report, 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. Important to note is that - in contrast with air traffic control - railway traffic control is much less prone to human errors leading to safety issues, as the signalling system ensures almost all safety aspects (Pellegrini and Rodriguez, 2013).

3.2 Data Envelopment Analysis

DEA 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, Seiford, and Zhu, 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 $\partial\widehat{\Psi}$ 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 Charnes, and Cooper (1984), takes scale differences into account when determining the production frontier $\partial\widehat{\Psi}$, 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 $\partial\widehat{\Psi}$ 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 Eeckhout, 1995) 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, Dyson, and Thanassoulis, 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 (1986a), 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, Seiford, and Zhu, 2011). The Banker and Morey (1986a) 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, Dyson, and Thanassoulis, 1991), or when the output 'quality of service' of municipalities is classified as 'good', 'normal' or 'bad' (Balaguer-Coll and Prior, 2009).

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

$$\widehat{\theta}_{VRS}^{l}(\boldsymbol{x},\boldsymbol{y}) =$$

$$min\left\{\theta > 0 \mid \theta\boldsymbol{x} \geq \sum_{i\in\bigcup_{f=1}^{l}K_f} \lambda_i\boldsymbol{x}_i\,;\ y \leq \sum_{i\in\bigcup_{f=1}^{l}K_f} \lambda_i\boldsymbol{y}_i\,;\ \sum_{i\in\bigcup_{f=1}^{l}K_f} \lambda_i = 1; \lambda_i \geq 0\,, i = 1\,,\dots,n\right\}, \qquad (1)$$

where the sample of n observations K = { 1, 2, …,n} is split into L subsets K_f ={ j | j ∈ K and level of the categorical variable = f} , and $K_i \cap K_j = \varnothing$, i ≠ j. In this formula, $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ 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 $\partial\widehat{\Psi}_{VRS}^{l}$ which, under the assumption of Variable Returns to Scale, tightly envelops all observations of level l and lower.

Bootstrapping the efficiency estimates

As the efficiency scores are based on the 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 VRS estimates. See Simar and Wilson (2008) for an overview and further technical details on this subject.

Now coming back to our DEA model with categorical variable, the integration of the variable implies that the convexity assumption is relaxed for this dimension of the model (Banker and Morey, 1986a). 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 technical details of the adapted algorithm are presented in the Appendix, to which we refer the interested reader.

3.3 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. 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).

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)[11]. 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 (BI) application code-named as the *CRIPTON BI 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[12]. 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 BI 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 electronic technology signal boxes, 14 relay-technology signal boxes 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:

[11] 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.

[12] 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.

Table 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 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.

Table 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

5. Results

5.1 DEA results

Table 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 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-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-16), while the Spearman coefficient equals 0.745 (p-value < 2.2e-16).

Taking a closer look at the gap between working week and weekend efficiency (see the histogram in figure 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.

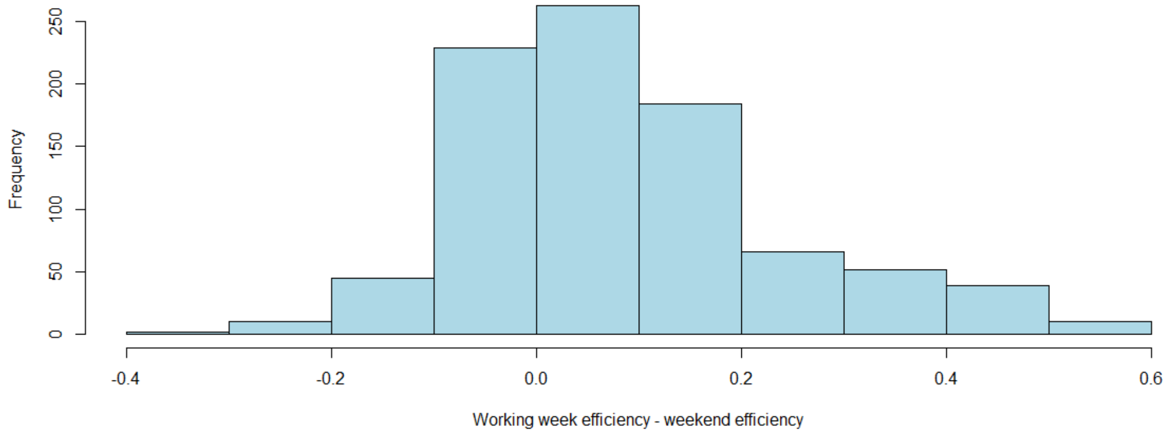

Figure 1. Histogram of observed efficiency differences between working week and weekend

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, Golany, and Li (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.

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[13] 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

[13] 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.

(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[14]. 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 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

[14] 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.

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 (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. InfraCost 2002, ORR report 2013, Button and Neiva 2013 and 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[15], as well as the semester dummies, and are therefore not reported. Including monthly dummies or a time trend provided similar results.

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

Table 4. Regression results on bias-corrected efficiency estimates

Variable	Working week (Mon-Fri)				Weekends (Sat-Sun)			
	OLS (robust SE)[a]	OLS (robust SE)[a]	trunc. bootstrap[b]	trunc. bootstrap[b]	OLS (robust SE)[a]	OLS (robust SE)[a]	trunc. bootstrap[b]	trunc. bootstrap[b]
	TR_AM	TR_AM_TC	TR_AM	TR_AM_TC	TR_AM	TR_AM_TC	TR_AM	TR_AM_TC
Constant	1.142***	0.435*	1.377***	0.414***	1.181***	0.630**	1.325***	0.617***
	(0.840, 1.444)	(-0.012, 0.881)	(1.222,1.507)	(0.247,0.582)	(0.911, 1.451)	(0.143, 1.118)	(1.216,1.393)	(0.491,0.737)
External decision-making (railway traffic characteristics)								
VAR	0.025	0.068	0.020	0.081***	0.039	0.052**	0.040***	0.056***
(variability)	(-0.130, 0.180)	(-0.046, 0.181)	(-0.042,0.083)	(0.035,0.127)	(-0.043, 0.122)	(0.005, 0.099)	(0.015,0.062)	(0.038,0.072)
DENS_SPAT	0.006**	0.005**	0.007***	0.004***	0.007	0.006	0.007***	0.006***
(spatial density)	(0.001, 0.011)	(0.001, 0.010)	(0.005,0.009)	(0.002,0.006)	(-0.003, 0.017)	(-0.002, 0.014)	(0.003,0.010)	(0.003,0.008)
DENS_TEMP	-0.059***	-0.034*	-0.072***	-0.025***	-0.021	-0.035	-0.019	-0.033***
(temporal density)	(-0.101, -0.016)	(-0.072, 0.004)	(-0.087,-0.052)	(-0.038,-0.011)	(-0.112, 0.070)	(-0.107, 0.036)	(-0.054,0.019)	(-0.056,-0.006)
TT_CHAR	0.015	-0.568	-0.919	-0.933	2.835	1.133	2.604***	1.020**
(timetable characteristics)	(-4.397, 4.426)	(-4.091, 2.955)	(-2.902,1.036)	(-2.327,0.501)	(-1.254, 6.924)	(-1.272, 3.538)	(1.169,3.852)	(0.098,1.890)
Internal decision-making (asset management policy)								
COMP_NET	-0.008	-0.054	-0.013	-0.057*	-0.049	-0.028	-0.047**	-0.026
(network complexity)	(-0.136, 0.120)	(-0.146, 0.038)	(-0.051,0.027)	(-0.090,-0.024)	(-0.163, 0.065)	(-0.104, 0.048)	(-0.080,-0.008)	(-0.056,0.004)
COMP_TRACK	-0.015***	-0.011***	-0.016***	-0.011***	-0.017***	-0.010***	-0.017***	-0.009***
(track complexity)	(-0.022, -0.008)	(-0.017, -0.005)	(-0.018,-0.013)	(-0.013,-0.009)	(-0.024, -0.009)	(-0.015, -0.004)	(-0.019,-0.014)	(-0.011,-0.008)
P_STATIONS	-0.005***	-0.004***	-0.007***	-0.005***	-0.006***	-0.005***	-0.007***	-0.005***
(proportion of stations)	(-0.006, -0.003)	(-0.005, -0.002)	(-0.008,-0.006)	(-0.006,-0.004)	(-0.008, -0.003)	(-0.006, -0.003)	(-0.008,-0.006)	(-0.005,-0.004)
WORK_DENS	0.018	0.012	0.018	0.009	-0.076**	-0.029	-0.084***	-0.036**
(infrastructure works)	(-0.027, 0.063)	(-0.026, 0.050)	(-0.011,0.050)	(-0.011,0.032)	(-0.134, -0.018)	(-0.079, 0.021)	(-0.118,-0.042)	(-0.060,-0.009)
Internal decision-making (traffic control policy)								
P_CLOSED		0.009***		0.012***		0.004***		0.005***
(closing times)		(0.005, 0.014)		(0.011,0.014)		(0.002, 0.006)		(0.004,0.006)
N_PERSONS		0.009***		0.011***		0.009***		0.009***
(team size)		(0.006, 0.013)		(0.009,0.013)		(0.004, 0.015)		(0.008,0.011)
KM_PERSON		0.034***		0.045***		0.062***		0.069***
(geographical centralisation)		(0.014, 0.054)		(0.038,0.051)		(0.047, 0.077)		(0.063,0.075)
R2 (adjusted R2)	0.424 (0.418)	0.645 (0.633)			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.

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' (UIC InfraCost study 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 (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 by the UIC InfraCost study (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[16] (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).

A first spin-off of our research is currently under development: a single-stage DEA model incorporating some of the significant environmental variables is being tested in the CRIPTON Business Intelligence tool. The model applies the Banker and Morey (1986b) approach, which allows for exogenous inputs and outputs. Also, imposing a limit on the importance of some of the environmental variables, custom constraints are added to the DEA linear programming problem (i.e. virtual weight restrictions, Wong and Beasley, 1990). In order to further improve the DEA model, our next research efforts will directed towards 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.

[16] 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.

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*. 8(4), 268-286.

Arellano, M. (1987). Computing Robust Standard Errors for Within-groups Estimators. *Oxford bulletin of Economics and Statistics*, 49(4), 431-434.

Balaguer-Coll, T., and Prior, D. (2009). Short-and long-term evaluation of efficiency and quality. An application to Spanish municipalities. *Applied Economics*, 41(23), 2991-3002.

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. (1986a). The use of categorical variables in data envelopment analysis. *Management science*, 32(12), 1613-1627.

Banker, R. D., & Morey, R. C. (1986b). Efficiency analysis for exogenously fixed inputs and outputs. *Operations Research*, *34*(4), 513-521.

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 (JTEP)*, 48(1), 65-80.

Cherchye, L., De Rock, B., 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.

Croissant, Y., Millo, G. (2008). Panel data econometrics in R: the plm package. *Journal of Statistical Software,* 27 (2), 1-43.

Cooper, W. W., Seiford, L. M., and Zhu, J. (2011). Data envelopment analysis: history, models, and interpretations. In Cooper, W. W., Seiford, L. M., and Zhu (Eds.) *Handbook on data envelopment analysis* (pp. 1-39). Springer US.

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* (Vol. 4). Springer.

EU-Commission. (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*.

EUROCONTROL Performance Review Commission. (2011). Econometric Cost-Efficiency Benchmarking of Air Navigation Service Providers. *EUROCONTROL, Brussels*.

EUROCONTROL Performance Review Commission. (2014). ATM Cost-Effectiveness (ACE) 2012 Benchmarking Report with 2013–2017 Outlook. *EUROCONTROL, Brussels*.

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*, 1-24.

Hansen, I. A., Wiggenraad, P. B. L., and Wolff, J. W. (2013). Performance analysis of railway infrastructure and operations. In *Proceedings of the 12th World Conference on Transport Research (WCTR 2013). Rio de Janeiro.*

Hayden-Smith, N. (2013). The Future of Signaller Workload Assessments in Automated World. In Dadashi, N., Scott, A., Wilson, JR, Mills, A. (Eds.) *Rail Human Factors: Supporting reliability, safety and cost reduction*. (pp. 419-426). CRC Press.

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

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*.

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. (2004). Assessing the efficient cost of sustaining Britain's rail network: Perspectives based on zonal comparisons. *Journal of Transport Economics and Policy*, 157-190.

Kneip, A., Simar, L., and Wilson, P. W. (2003). *Asymptotics for DEA estimators in nonparametric frontier models* (Vol. 317). Université Catholique de Louvain, Louvain-la-Neuve, Discussion paper 0317.

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

Merkert, R., and Nash, C. (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.

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*. 2nd edition, 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. In Fried, H. O., Lovell, C. K., & Schmidt, S. S. (Eds.) *The measurement of productive efficiency and productivity growth*, 421-521. Oxford University Press.

Simar, L., and Wilson, P. W. (2011). Two-stage DEA: caveat emptor. *Journal of Productivity Analysis*, 36(2), 205-218.

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. (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., 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.

Tulkens, H., and Vanden Eeckaut, P. (1995). Non-parametric efficiency, progress and regress measures for panel data: Methodological aspects. *European Journal of Operational Research*, 80(3), 474-499.

UIC Infrastructure Commission. (2002). InfraCost - The Cost of Railway Infrastructure. *Final Report, Paris*.

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

Wheat, P., and Smith, A. (2008). Assessing the marginal infrastructure maintenance wear and tear costs for Britain's railway network. *Journal of Transport Economics and Policy*, 189-224.

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

Appendix: DEA subsample bootstrap algorithm, adapted for categorical variable

To obtain bias-corrected efficiency VRS estimates, we apply the subsample bootstrapping algorithm proposed in the literature. See Simar and Wilson (2008) for an overview and further technical details on this subject.

The purpose of this Appendix is to present a subsample bootstrap algorithm applicable to the Banker and Morey (1986a) DEA models with a categorical variable.

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

$$\hat{\theta}_{VRS}(\boldsymbol{x},\boldsymbol{y}) = \\ min\left\{\theta > 0 \mid \theta\boldsymbol{x} \geq \sum_{i=1}^{n} \lambda_i \boldsymbol{x}_i \,;\, y \leq \sum_{i=1}^{n} \lambda_i \boldsymbol{y}_i ; \sum_{i=1}^{n} \lambda_i = 1; \lambda_i \;\geq 0\,, i = 1\,, \ldots, n\right\}. \tag{2}$$

Here, $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ 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) to construct the convex and free-disposal hull $\hat{\Psi}_{VRS}$, tightly enveloping these observations.

The idea behind the bootstrap procedures is to approximate the unknown sampling distribution of $\hat{\theta}_{VRS}(\boldsymbol{x},\boldsymbol{y}) - \theta(\boldsymbol{x},\boldsymbol{y})$ through the empirical distribution of $\hat{\theta}^*_{VRS}(\boldsymbol{x},\boldsymbol{y}) - \hat{\theta}_{VRS}(\boldsymbol{x},\boldsymbol{y})$, in which $\hat{\theta}^*_{VRS}(\boldsymbol{x},\boldsymbol{y})$ 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 $\hat{\theta}^*_{VRS}(\boldsymbol{x},\boldsymbol{y})$, is known to be inconsistent[17]. Two solutions providing consistent inference have been proposed by Kneip, Simar, and Wilson (2003): 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 $\hat{\theta}^*_{VRS,m,b}(\boldsymbol{x},\boldsymbol{y})$ is calculated with the bootstrap sample $S^*_{m,b} = \{(\boldsymbol{x}_i^{*,b}, \boldsymbol{y}_i^{*,b}), i = 1, \ldots, m)\}$ determining the bootstrap production possibility set:

$$\hat{\theta}^*_{VRS,m,b}(\boldsymbol{x},\boldsymbol{y}) \; = \\ min\left\{\theta > 0 \mid \theta\boldsymbol{x} \geq \sum_{i=1}^{m} \lambda_i \boldsymbol{x}_i^{*,b} \,;\, y \leq \sum_{i=1}^{m} \lambda_i \boldsymbol{y}_i^{*,b} ; \sum_{i=1}^{m} \lambda_i = 1; \lambda_i \;\geq 0\,, i = 1\,, \ldots, m\right\}. \tag{3}$$

For the subsample bootstrap, Kneip, Simar, and Wilson (2003) have proven that as the number of bootstrap iterations B → ∞, the empirical distribution of $m^{\overline{(N+M+1)}^{2}}\left(\hat{\theta}^*_{VRS,m,b}(\boldsymbol{x},\boldsymbol{y}) - \hat{\theta}_{VRS}(\boldsymbol{x},\boldsymbol{y})\right)$ approximates the unknown sampling distribution of $n^{\overline{(N+M+1)}^{2}}\left(\hat{\theta}_{VRS}(\boldsymbol{x},\boldsymbol{y}) - \theta(\boldsymbol{x},\boldsymbol{y})\right)$. 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 formula 1), the integration of the variable implies that the convexity assumption is relaxed for this dimension of the model (Banker

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

and Morey, 1986a). 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.

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 ≤ l.

Thus, after calculating the efficiency estimates $\widehat{\theta}^{l}_{VRS}(\boldsymbol{x},\boldsymbol{y})$ with formula (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 ≤ 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 $\hat{\theta}^{*,l}_{VRS,m_l,b}(\boldsymbol{x},\boldsymbol{y})$ using the bootstrap pseudo-sample $S^{*}_{m_l,b}$ from the previous step, and applying formula (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, …, B.
4. The resulting B bootstrap values can then be used to estimate the bias for each observation:

$$\widehat{BIAS}_B\left(\hat{\theta}^{l}_{VRS}(\boldsymbol{x},\boldsymbol{y})\right) = \left(\frac{m_l}{n_l}\right)^{\frac{2}{(N+M+1)}} \left[\frac{1}{B}\sum_{b=1}^{B} \hat{\theta}^{*,l}_{VRS,m_l,b}(\boldsymbol{x},\boldsymbol{y}) - \hat{\theta}^{l}_{VRS}(\boldsymbol{x},\boldsymbol{y})\right], \qquad (4)$$

   with the factor $\left(\frac{m_l}{n_l}\right)^{\frac{2}{(N+M+1)}}$ 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:

$$\hat{\hat{\theta}}^{l}_{VRS} = \hat{\theta}^{l}_{VRS}(\boldsymbol{x},\boldsymbol{y}) - \widehat{BIAS}_B\left(\hat{\theta}^{l}_{VRS}(\boldsymbol{x},\boldsymbol{y})\right). \qquad (5)$$

We wrote the algorithm elaborated above in R, using the functions of the FEAR 2.0 package (Wilson, 2008). Bootstrap calculations were performed applying 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).