ECARES

Performance Analysis with Unobserved Inputs: An Application to Endogenous Automation in Railway Traffic Management

Laurens Cherchye
Department of Economics, KULeuven

Bram De Rock
ECARES, Université libre de Bruxelles and KU Leuven

Dieter Saelen
Department of Economics, KULeuven
National Bank of Belgium

Marijn Verschelde
Department of Economics & Quantitative Methods, IESEG School of Management, Lille
Department of Economics, KULeuven

Bart Roets
Faculty of Economics and Business Administration, Ghent University
On Track Lab research center and Performance Data Division, Infrabel

March 2022

ECARES working paper 2022-06

ECARES
ULB - CP 114/04
50, F.D. Roosevelt Ave., B-1050 Brussels BELGIUM
www.ecares.org

Performance analysis with unobserved inputs: An application to endogenous automation in railway traffic management*

Laurens Cherchye[†] Bram De Rock[‡] Dieter Saelens[§] Marijn Verschelde[¶]
Bart Roets[‖]

Abstract

Performance analytics are commonly used in managerial decision making, but are vulnerable to an omitted variable bias issue when there is incomplete information on the used production factors. In this paper, we relax the standard assumption in productive efficiency analysis that all input quantities are observed, and we propose a nonparametric methodology for cost inefficiency measurement that accounts for the presence of unobserved inputs. Our main contribution is that we bridge the OR/MS and the economic literature by addressing the general critique of Stigler (1976) on the concept of inefficiency (Leibenstein, 1966), which states that found inefficiencies reflect unobserved inputs rather than waste. Our methodology explicitly differentiates between cost inefficiency (i.e. waste; deviations from optimizing behavior) and unobserved input usage (i.e. optimally chosen input factors that are unobserved to the empirical analyst). We apply our novel method to a purpose-built dataset on Belgian railway traffic management control rooms. Our findings show the existence of meaningful inefficiencies that cannot be attributed to use of unobserved inputs or environmental factors. In addition, we document how the omitted variable bias impacts cost efficiencies of individual observations in a dissimilar way in case the use of unobserved inputs is not controlled for.

Keywords: efficiency measurement, unobserved heterogeneity, omitted variable

*We thank Peter Bogetoft and Tom Van Puyenbroeck for insightful discussion and feedback. We are also grateful to the seminar audience in Leuven and to participants of the NAPW conference 2020 and 2021 in Miami, the Informs Annual Meeting 2021 and the 31st European Conference on Operational Research (EURO 2021).

[†]Department of Economics, University of Leuven. E-mail: laurens.cherchye@kuleuven.be.

[‡]Ecares, Université Libre de Bruxelles, and Department of Economics, University of Leuven. E-mail: bram.de.rock@ulb.be.

[§]Department of Economics, University of Leuven and National Bank of Belgium. E-mail: dieter.saelens@kuleuven.be.

[¶]Department of Economics and Quantitative Methods, IÉSEG School of Management, Lille Économie Management (UMR-CNRS 9221) and Department of Economics, University of Leuven. E-mail: m.verschelde@ieseg.fr.

[‖]On Track Lab research center and Performance Data Division, Infrabel and Faculty of Economics and Business Administration, Ghent University. E-mail: bart.roets@infrabel.be.

bias, cost minimization, railways.
JEL: D24, L23, L25

Disclaimer: The views expressed in this article are those of the authors and should not be interpreted as reflecting the views of Infrabel.

1 Introduction

Performance measures and their corresponding prescriptive analytics are of increasing importance for managerial decision making.[1] Therefore, the measurement of productive performance has been the focus of a well-established branch of the operations research and management science (OR/MS) literature and the economic literature. Performance measurement at any level is however intricate as the empirical analyst is confronted with many empirical challenges and usually has incomplete information about the production process. The latter resulted in a significant divide between the economic literature and the OR/MS literature. Within the OR/MS field, a growing efficiency measurement literature builds on the seminal works of a.o. Leibenstein (1966) that clarify the concept of inefficiency as a representation of suboptimal conduct of the decision making unit (i.e. waste). After the general critique of Stigler (1976), which states that found inefficiencies reflect unobserved (to the analyst) inputs, economists largely rejected the idea that performance heterogeneity could be driven by suboptimal behavior and waste. The economic literature on production analysis attributes performance heterogeneity to unobserved inputs or exogenous shocks, that is, unobserved heterogeneity within a framework of optimizing economic agents.

Empirical setting. The increasing prevalence of intangible inputs, automation and digitization that characterize many contemporaneous business environments warrants a unifying framework that allows for both inefficiency and unobserved inputs. Indeed, the intangible nature of (technological) inputs challenges the standard assumption of perfect knowledge of input costs that is often maintained within the OR/MS literature.[2] As such, existing methods for performance benchmarking may be subject to an omitted variable bias when studying digitizing settings characterized by intangible input usage (i.e. technology, management quality, automated decision support, etc.). Railway traffic management in Belgium is a prime example of a setting with a need for performance benchmarking while controlling for increasing uses of automation and digitization. Railway traffic management decisions concerning work design and automation are made

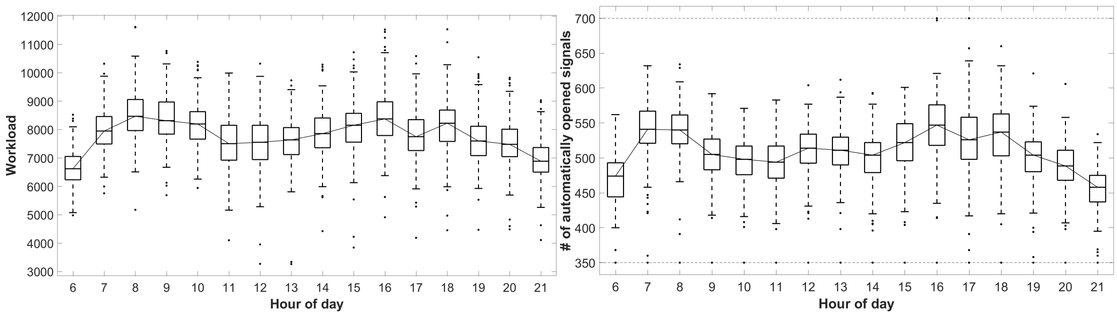
[1] See e.g. Abernethy et al. (2021) and references therein for empirical evidence of the influence of performance measurement on strategical decision making and firm performance. Bryniolfsson and McElheran (2016); Wu et al. (2020) show evidence of the increasing use of data analytics for decision making and how the use of data-driven decision making relates to firm productivity and innovation.

[2] See e.g. Bloom and Van Reenen (2007); Saunders and Brynjolfsson (2016); Nagle (2019); Brynjolfsson et al. (2021) for discussions of the measurement of intangible assets and its implications for performance measurement.

within the context of digitized traffic management control rooms. These traffic management control rooms are an ideal testing ground to analyze performance and optimal work organization via our advocated framework.

At Infrabel, the Belgian railway infrastructure company, an automated route-setting (ARS, (Pachl, 2009)) device has been implemented which allows staff to autonomously choose when and to what extent to automate movement decisions. The endogenous nature of this traffic automation input, in combination with its intangible features, will render existing performance benchmarking methods (such as Data Envelopment Analysis (DEA)) vulnerable to an omitted variable bias when these unobserved inputs are not controlled for. This mechanism is made explicit in Figure 1 where the left panel shows hourly variation in workload levels for a specific control room, while the right panel exhibits hourly levels of automation and use of the ARS.[3] There exists a clear link between workload and the level of traffic automation with staff making increasing use of automation during peak hours while reducing automation levels amid off-peak hours. Existing methods will overlook the fact that part of the variation in workload is unrelated to staffing levels but instead is explained by the level of automation, leading to a biased efficiency measurement. More precisely, since they neglect unobserved input usage existing benchmarking methods will underestimate the aggregate/total quantity of used input, thereby creating artificially large inefficiency differentials across observations.

Figure 1: Control room 1 – Boxplots of workload and automation



Note: Workload represents the time-weighted sum of all operational traffic management actions (such as train movements, merging or splitting of trains, changing track lines) carried out by the team of traffic control at work during the hour of evaluation.

Methodological contribution. The present paper proposes a method for efficiency measurement that controls for omitted variable bias originating from endogenous inputs which are unobserved by the empirical analyst. This allows us to account for the endogeneity issue in efficiency analysis that is caused by correlation between unobserved heterogeneity and observed input usage (see Marschak and Andrews (1944);

[3]Section 2 explains in detail how we computed the different variables in our analysis.

Olley and Pakes (1996)). Our proposed framework is embedded within the nonparametric production literature. Nonparametric approaches impose no a priori functional form relationships. This is especially convenient for our empirical application, as it would be particularly difficult to convincingly justify a specific parametric specification for the complex setting at hand.[4]

More specifically, we extend the nonparametric framework for production analysis of Cherchye et al. (2021) by explicitly allowing for wasteful/inefficient behavior. Cherchye et al. (2021) proposed a nonparametric methodology to estimate productivity (as exogeneous shock or unobserved input) and output elasticities under endogeneity. As such, they provide a nonparametric alternative for the standard production function approaches to deal with the simultaneity bias that appeared in the economic literature (see, for example, Olley and Pakes (1996); Ackerberg et al. (2015)). We structure unobserved input costs to be observed and/or predictable by the firm, but unobserved by the empirical analyst. Unobserved input costs can relate to all aspects of productivity (Syverson, 2011) that do not reflect waste (as we capture the latter by the efficiency term). This includes intangibles, managerial quality, anticipated exogenous shocks, automated decision support, etc..

In the OR/MS literature, efficiency measurement involves estimation of suboptimal behavior (i.e. inefficiency), which is ruled out by assumption in the economic literature on productivity. Since the pioneering works on the Data Envelopment Analysis (DEA; Charnes et al. (1978)) and Free Disposal Hull (FDH; De Prins et al. (1984)) efficiency estimators, numerous applications within the operations research literature have maintained the implicit assumption that the analyst has full knowledge of all factor quantities underlying production. By extending Cherchye et al. (2021), we can relax this assumption and identify both efficiency and unobserved inputs under endogeneity. We show that our approach has empirical bite and adequately controls for unobserved inputs and potential correlation between unobservables and observed input usage by means of a Monte Carlo simulation.

Several approaches are available in the DEA literature to account for factors different from the observed inputs and outputs that affect productive efficiency performance. Yet, none of the existing methods is sufficient for analysing the current setting. First, the endogenous nature of automation use will violate the assumption of separability, thus

[4]In the context of nonparametric efficiency analysis, Cordero et al. (2015) show via a Monte Carlo exercise that dependency of input choice on efficiency can imply severely biased estimates of efficiency. Simar et al. (2016) deal with unobserved heterogeneity in environmental variables that influence the production function by structuring the potential influence of unobserved environmental factors. Santín and Sicilia (2017) and Cazals et al. (2016) propose an instrumental variable approach to allow for dependency of input choice on efficiency. In the current paper, we allow for dependency of input choices on anticipated unobserved heterogeneity. We exclude dependency of the frontier on unanticipated efficiency (e.g. caused by reversed causality), in line with the idea that inefficiency reflects waste that is unobserved by the firm when making input choice decisions.

rendering invalid the outcomes of a classical two stage approach (Simar and Wilson, 2007, 2011)). Second, the intangible nature of automation makes it nearly impossible to obtain a perfect proxy of unobserved inputs required to implement the (conditional) order-m approach of Daraio and Simar (2005, 2007) (which relies on the assumption of perfect knowledge of input(s) (costs)). This again illustrates a need to develop efficiency measures explicitly recognizing that some inputs may be hard to measure, which forms exactly the core aim of the current paper. We will however show that our proposed method is complementary to the (conditional) order-m method. Lastly, also the method of Simar et al. (2016) for dealing with unobserved heterogeneity and endogeneity is not directly useful in our application setting, as unobserved production heterogeneity takes the form of an endogenous input rather than an exogenous environmental factor that falls beyond the control of the firms.

Empirical contribution. Our application to railway traffic management demonstrates that the inclusion of unobserved inputs does not imply inexistence of inefficiency. We find empirical support for the existence of inefficiency, even when controlling for optimally chosen inputs that are unobserved (or only partly observed) to the empirical analyst. In particular, while we report lower rates of inefficiency when accounting for unobserved input usage, we still find persistent inefficiencies of over 20% that cannot be ascribed to inattention to an unobserved input. Differently stated, we show that performance heterogeneity can reflect both waste and unobserved inputs. We thus find no supportive evidence for the general critique of Stigler (1976) that the concept of inefficiency (Leibenstein, 1966) merely reflects unobserved inputs.

Next, we illustrate the value of disaggregated data to unravel the micro-dynamics of the omitted variable bias and cost efficiency heterogeneity. We report that the effect of accounting for unobserved inputs on the efficiency estimates is inversely related to the level of traffic automation. Our cost efficiency estimates of individual observations are thus distorted in a dissimilar way in case unobserved inputs are not controlled for. Moreover, our results show a clear link between cost inefficiencies and hour-of-the-day, highlighting that the scheduling system with three non-overlapping eight hour shifts requires customisation to better fit the hourly variation and peaks in traffic controllers' workload.[5]

More generally, we show the value of automation to complement employees in digitized environments. Our empirical application indicates that traffic automation serves as a way to mitigate the impact of unforeseen events on the workload of control room operators. We thus present empirical evidence suggesting human-automation complementarity. Prevention of over-workload as measured as super-efficiency is important in the context of digitized environments from an employee well-being, quality of service

[5] We refer to Aksin et al. (2007) for an overview of operations management and staff scheduling in the related context of call centers and to Campbell (1999) as an early reference on the benefits of cross-trained staff in the context of the staff scheduling problem.

and safety perspective.[6] In the context of our application, managers continuously make decisions on the allocation of subzones to the different operators in order to prevent over-workload. By including both efficiency and automation in a single analytical framework, as we advocate, managers can better customise their decision making to the needs of the situation.

Outline. This paper is organized as follows. Section 2 presents a detailed discussion of the production environment of railway management control rooms in Belgium. This will motivate the use of our novel methodology for this application setting. Section 3 presents our theoretical framework for efficiency measurement in the presence of unobserved inputs. We also introduce our cost efficiency measure that we use to quantify the degree of waste in production, and we discuss a number extensions that will be useful when bringing our methodology to data in our empirical application. Section 4 presents our main empirical findings, and Section 5 highlights the associated managerial implications. Section 6 concludes.

2 Railway traffic control in Belgium

Control rooms in Belgian railways are a prime example of a production setting within a digitized environment. As firms use more frequently digital components, the unobserved production factors are increasing in importance. Therefore, we expect an omitted variable bias when not adequately controlling for both observed and unobserved input costs in production settings with intangible production factors. To study traffic controllers' performance, we built a data structure that includes next to input-output data on the control room-hour level, information on exogenous environmental variables such as lagged delays, density of the railway network and the interrelation with safety controllers. As such, we are the first to disentangle influences from (i) unobserved inputs (i.e. ARS, intangibles, management quality, etc.), (ii) cost inefficiency (i.e. waste due to inflexible scheduling, etc.), and (iii) environmental factors (i.e. unanticipated workload due to delays, safety considerations, etc.). We estimate hourly cost efficiency and unobserved input shares over 12 months in 2018-2019 within 9 control rooms.

The main task of these control rooms is the real-time coordination of all railway traffic happening in Belgium. Following Roets, Verschelde, and Christiaens (2018, p.228), we define railway traffic management 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). In our application we will focus on the operational aspects of traffic management (i.e. signalling activities and traffic management) as it are these aspects that are partly

[6] In the context of restaurants, Tan and Netessine (2014) show causal evidence that high workload can lower service quality. See Roets et al. (2018) for a discussion on the relation between over-workload and super-efficiency in our application context.

automated in an endogenous fashion by ARS. We will, however, consider safety interventions (that are exogenous to traffic operators by work design) as an environmental variable in our analysis.

We focus on *cost efficiency as representing waste*. This concerns predominantly overpresence of traffic operators due to, among other things, suboptimal or inflexible work design. Output quantity and input prices (i.e. wages) are exogenous to the control rooms, while the local management of the control rooms holds a certain degree of flexibility in organizing the optimal alignment of resources with these outputs.[7] In this setting questions regarding the optimal organization of work arise naturally. Ideally, one would like to tailor working conditions in function of specific hourly needs. This would however impose unrealistic demands on both management and personnel. Further, several important requirements complicate the adjustment of working conditions in these control rooms. Traffic management control rooms monitor the railway traffic in a 24/7 fashion with a work schedule consisting of three non-overlapping eight hour shifts (Topcu et al., 2019). Staffing levels generally remain constant within the shift, although the volume of railway traffic may (strongly) vary across adjacent hours. This requires that throughout the shift a sufficient amount of staff is being deployed to manage fluctuations in railway traffic at all times, including peak hours. As a consequence of the constraints of the scheduling system we expect to recover manifest differences in efficiency between hours. Taking these different considerations into account we believe that railway traffic management provides an excellent case to consider cost efficiency when some input costs are unobserved.

We include in our model both observed and unobserved input costs. Concerning the *observed input costs*, each traffic management control room is organised according to multiple controller and supervisor roles.[8] Out of these, Traffic Controllers (TCs), who are responsible for operational traffic management (e.g. opening/closing signals, merging or splitting trains, changing train tracks, etc.), form the object of our interest. More precisely, the empirical application differentiates between three types of TCs based on an internal grade system, with TCs of a higher grade receiving higher compensation. We consider hourly wages for the TCs, which differ across grades but are otherwise constant. For confidentiality reasons, we do not report these wages.

Unobserved inputs relate to inputs that are observed or predictable to the control room management, but are unobservable for the empirical analyst. Its main components in the empirical setting of focus are the use of ARS, management quality and the value of intangibles such as customized software. We collected data to put structure on the value of the unobserved input costs within subgroups of our dataset. In particular, we

[7] Output is exogenous to the control room as the Belgian railway transportation company NMBS, which is independent from Infrabel, is responsible for the drafting of the train schedule.

[8] See Topcu et al. (2019) for an in-depth discussion of the inner working of the traffic management control rooms.

use the prevalence of hourly automated signal openings, which we define as the number of automatically opened signals. This measure serves as a reasonable proxy for the use of the ARS system available in all traffic management control rooms. Specifically, ARS provides TCs with a sort of autopilot mode that performs automated movement decisions when a train approaches a signal. Note that TCs can choose on the spot whether to manually open a signal or to use the ARS for an automatic signal opening.[9] As such, a higher amount of automated signal openings implies a higher usage of the ARS and thus a higher level of traffic automation. We assume, within each control room, that all unobserved inputs other than ARS usage are constant within a week. Although this may seem rather strict at first, it links up naturally with the planning of staffing schedules. At Infrabel, staffing decisions are agreed upon on a monthly basis (which assumes constant levels of, for example, management quality within a given month). However, deviations from these standard monthly staffing levels might be necessary to process for example urgent track repair works (which sometimes requires extra personnel to monitor railway traffic) or to accommodate sickness (i.e. replacement by a TC of a different grade). To account for this, we only impose these other factors to remain constant on a weekly basis. We then use the strict monotonic relation between the hourly number of automated signal openings and the unobserved input costs. More precisely, within every week of data we impose the ordering of this variable onto the unobserved input costs.

Further, we include *three environmental variables* that interfere with the work of TCs. First, previous train delays potentially disturb a fluent and safe traffic flow and demand additional consideration from the TCs. We consider the average delay in seconds (aggregated over all TCs) measured during the last 15 minutes of the previous hour.[10] Second, we account for fluctuations in traffic density, which peaks during the morning and late afternoon rush hours.[11] Third, safety personnel in charge of monitoring traffic safety can intervene with TCs operations for safety reasons. To account for this we include as an extra environmental variable the time-weighted sum of operational traffic management

[9]Remark that it is not an option for TCs to simply rely on ARS all the time. Even contemporary ARS systems are to some extent rudimentary (Balfe et al., 2015). For example, the ARS may wrongly give priority when multiple trains attempt to simultaneously pass through bottlenecks, especially as the network becomes more and more congested. Moreover, automation systems may lack the necessary flexibility needed to appropriately respond to unforeseen events (which may occur frequently in open loop systems such as railway traffic control). As such, TCs need to constantly trade-off monitoring (and possibly adjusting) ARS behaviour (which may increase mental workload and TC fatigue) against conducting tasks manually themselves.

[10]Arguably, delays can be both endogenous (e.g. human error/miscommunication within the traffic management control rooms) and exogenous (e.g. technical issues, infrastructure or component breakdowns, accidents, snowfall, human error outside the traffic management control room) in nature. Consultation with Infrabel experts made clear that the endogenous part of delays is negligible in the current setting. We thus consider the underlying causes for delays to be exogenous in nature. This way we obtain a benefit-of-the-doubt analysis that estimates cost efficiencies filtered for the possible influence of delays. Following expert advise, average delays are calculated by averaging the difference in seconds between planned and realized signal passing, thus comprising both trains ahead and behind of schedule.

[11]The density variable measures the ratio of the number of train passings at certain key signals over the number of key signals.

tasks carried out by the safety personnel of the control room in the hour of evaluation.

Table 1 provides an overview of the inputs, output and environmental variables that we consider in the empirical model. Our inputs measure the number of TCs of a specific grade working at the control room during the hour of evaluation. Next, the output variable outputTC represents the time-weighted sum of all operational traffic management actions (such as train movements, merging or splitting of trains, changing track lines) carried out by the team of TCs at work during the hour of evaluation. As in Roets et al. (2018), the weight of each action is determined by the action's standard execution time in seconds, as judged by an expert panel. Finally, we obtain a sample of 28.778 hourly observations across 9 traffic management control rooms over the period June 2018 to May 2019.[12] Table 2 and Table 3 present, respectively, descriptive statistics and correlations for the variables used in our empirical model. Both output and the number of TCs correlate strongly with the number of automated signal openings, hinting at a potential omitted variable bias.

Table 1: Overview of input, output and environmental variables.

Type	Name	Definition
Input	TC_N3	Number of TCs (lowest grade)
Input	TC_N2	Number of TCs (medium grade)
Input	TC_N1	Number of TCs (highest grade)
Output	outputTC	Weighted number of operational traffic management actions of TCs (by standard times)
Env. var.	Delay	Mean delay time (in seconds) during the last 15 minutes of the previous hour (average over all TCs)
Env. var.	Density	Mean traffic density (average over all TCs)
Env. var.	Workload_SC	Weighted number of operational traffic management actions of safety personnel (by standard times)

[12] For three control rooms we consider a shorter time window, due to data irregularities.

Table 2: Descriptive statistics

	Mean	Median	St. Dev.	Min	Max
TC_N3	5.49	5	1.84	0	10
TC_N2	1.75	2	1.27	0	7
TC_N1	0.15	0	0.39	0	3
outputTC	6308	6376	1925	1266	17312
Delay	143	115	126	-645	1769
Density	667	623	245	57	2035
WorkloadSC	940	711	816	0	6124
# Aut. signal openings	322	290	116	23	904

Table 3: Correlation table

	TC	OutputTC	Delay	Density	WorkloadSC	# Aut. signal openings
TC	1					
OutputTC	.497	1				
Delay	.004	.166	1			
Density	-.176	.374	.044	1		
WorkloadSC	.254	.252	.109	-.134	1	
# Aut. signal openings	.514	.776	.006	.461	.123	1

Note: TC sums traffic controllers of all grades.

3 Efficiency measurement methodology

Our novel framework is embedded within the structural, nonparametric production literature and extends recent insights of Cherchye et al. (2021) by introducing measures of cost efficiency to recover unobserved heterogeneity of cost-minimizing firms. After introducing some necessary notation we provide a brief overview of the model used in Cherchye et al. (2021). This will pave the way for introducing our cost efficiency criterion. Finally, we discuss some extensions to the basic model.

We refer to Appendix A.2 for a Monte Carlo simulation highlighting the finite sample performance of our estimator. Our Monte Carlo simulation shows the proper working of our estimators for noisy production data and settings with biased technological change under endogeneity. In addition, our empirical results in Section 4 show that our approach adequately deals with the important omitted variable issue related to endogenous automation in our application.

3.1 Preliminaries

We start our methodology from a set T of hourly observations of control rooms, which we consider as Decision Making Units (DMUs) that operate under a common production technology. These DMUs use several inputs to produce a single output (i.e. weighted number of operational traffic management actions OutputTC), denoted by Q. For our DMUs, we observe the active operators of different grades (i.e. TC_N1, TC_N2 and TC_N3). We denote the observed inputs by the 3-dimensional vector $\mathbf{X} \in \mathbb{R}^3_+$ and the observed input prices by $\mathbf{W} \in \mathbb{R}^3_{++}$. We follow standard practice in the literature (see Olley and Pakes (1996); Ackerberg et al. (2015)) by including unobserved inputs as a one-dimensional aggregate. In particular, we treat unobserved inputs as a Hicksian aggregate under the assumption of a common unobserved input price within the control rooms.[13] We denote this with $\Omega \in \mathbb{R}^+_0$. Our method does not require information regarding output prices.

Building upon the work of Varian (1984), the contribution of Cherchye et al. (2021) develops a nonparametric model to test consistency of firm demand and supply data with cost minimization in the presence of unobserved heterogeneity. Specifically, the model assumes a common production function F (strictly monotonic, continuous and quasi-concave in $(\mathbf{X}, \Omega)$) that defines

$$Q = F(\mathbf{X}, \Omega).$$

We assume that F exhibits constant returns to scale (CRS) w.r.t. observed inputs $\mathbf{X}$, for a given amount of the unobserved input Ω:

Axiom 1 (CRS on observed inputs)**.**
$F(k\mathbf{X}, \Omega) = k * F(\mathbf{X}, \Omega)$ where $k = \mathbb{R}^+_0$.

Whether or not to impose CRS on both the observed inputs $\mathbf{X}$ and/or unobserved inputs Ω is somewhat application-specific and will depend on the nature of the unobserved input. For example, in the current application the unobserved input measures, among other things, the value of the ARS, which can be used in a non-rival way. Thus, a rescaling of the unobserved input is not warranted for the empirical application of focus (e.g. we do not require that doubling computers doubles output levels). While alternative returns to scale assumptions can be easily incorporated, consultation with Infrabel experts confirmed that the above CRS assumption is realistic for our application setting.

Next, unobserved Ω can represent either an external factor beyond the DMUs control or an optimally chosen latent input factor, implying two different optimization problems. Both interpretations are shown to be empirically equivalent by Cherchye et al. (2021),

[13] See Hicks (1946). In principle, it is possible to extend our framework to include multiple unobserved inputs, but this requires additional structure and goes beyond the scope of the current paper. In the setting of focus, we believe that our one-dimensional aggregate realistically covers the total variation of unobserved input use, especially when additional structure is included, as discussed in Section 3.3.

giving rise to exactly the same nonparametric testable implications. As in our empirical setting the unobserved input costs can be considered to be endogenous, we here only consider the optimization problem of the second case. More specifically, we consider each DMU to solve the following optimization problem:

$$\min_{\mathbf{X},\Omega} \mathbf{W}\mathbf{X} + \Omega \text{ s.t. } F(\mathbf{X},\Omega) \geq Q, \tag{OP}$$

meaning that each DMU chooses observed input values $\mathbf{X}$ and unobserved input value Ω to produce the output level Q at minimum cost.

Basically, our proposed method checks cost efficiency for all DMUs simultaneously by verifying whether their observed behavior can be rationalized as cost minimizing (in the tradition of Varian (1984) and Banker and Maindiratta (1988)). In particular, our analysis starts from a set of observed data within a control room setting $S = \{Q_t, \mathbf{X_t}, \mathbf{W_t} \mid t \in T\}$, containing information on a set T of hourly input-output observations (treated as DMUs). The unknowns for the analyst are the functional form of F which is specific for each control room, and the level of unobserved inputs Ω which is observation-specific. We verify whether the observed behavior in S is consistent with cost minimizing production by checking whether there exists at least one specification for F and unobserved input values Ω_t ($t \in T$) that represent the observed behavior as solving problem OP. If there exist such F and Ω_t, the dataset S is said to be OP-rationalizable, implying that the observed behavior in S can be labelled as cost efficient. This is summarized the following definition.

Definition 1. The dataset $S = \{Q_t, \mathbf{W_t}, \mathbf{X_t}\}_{t\in T}$ is OP-rationalizable if there exist numbers $\Omega_t \in \mathbb{R}_0^+$ and a production function $F : \mathbb{R}_+^{N+1} \to \mathbb{R}_+$ such that for each DMU $t \in T$,

$$(\mathbf{X_t}, \Omega_t) \in \arg \min_{\mathbf{X},\Omega} \mathbf{W_t}\mathbf{X} + \Omega \text{ s.t. } F(\mathbf{X},\Omega) \geq Q_t.$$

Cherchye et al. (2021) derived the following characterization of OP-rationalizability, defining a nonparametric testable condition for data consistency with cost minimization:

Proposition 1. *Let $S = \{Q_t, \mathbf{W_t}, \mathbf{X_t}\}_{t\in T}$ be a given dataset. The following statements are equivalent:*

1) The dataset S is OP-rationalizable.

2) There exist $\Omega_t \in \mathbb{R}_0^+$ that satisfy, for all $t, s \in T$, the inequalities

$$\mathbf{W_t}\mathbf{X_t} + \Omega_t \leq \mathbf{W_t}\left[\left(\frac{Q_t}{Q_s}\right)\mathbf{X_s}\right] + \Omega_s.$$

The testable requirement in statement (2) of this result naturally extends Varian (1984)'s Weak Axiom of Cost Minimization (=WACM) towards settings characterized by unobserved inputs. In words, it verifies whether there exist unobserved input levels Ω_t such that for every observation t, the output level Q_t is effectively produced at the lowest possible cost w.r.t. the rescaled input bundles of the other observations s. This condition has an intuitive interpretation. Specifically, the CRS assumption in Axiom 1 ensures, for given Ω_s, that the rescaled input bundles $(\left(\frac{Q_t}{Q_s}\right)\mathbf{X_s})$ can produce the output level Q_t.[14] The right-hand-side of the inequality in statement (2) then shows the cost of using this rescaled input bundle at the input prices $\mathbf{W_t}$ that apply to observation t. Cost minimization requires that, for the given prices, the actual cost at observation t (using the input bundle $(\mathbf{X_t}, \Omega_t)$ to produce Q_t) cannot exceed this cost of any rescaled input bundle.

Appendix A.1 provides numerical examples highlighting the empirical content of the cost minimization condition in Proposition 1. Specifically, we establish that our empirical restrictions for cost minimizing production (under the presence of unobserved inputs) can be rejected in a minimalistic setting considering only two DMUs and two observed inputs. In general, the empirical bite of the testable requirement will increase with the number of observations and observed inputs.

3.2 Cost efficiency

Checking the "sharp" cost minimization requirement in statement (2) of Proposition 1 yields a binary outcome: the observed behavior in a given control room (represented by the dataset S) either satisfies the condition or not. OP-rationalizability will be violated as soon as there exists a single DMU that behaves inefficient. This limits the empirical usefulness of the testable requirement in its strict form. For example, our empirical application in Section 4 will show that there are many cost inefficient observations for all control rooms that we evaluate. This falls in line with the claim of Banker and Maindiratta (1988), who argue that in many empirical environments it is very likely that at least some DMUs will be characterized by inefficient production processes. Moreover, when violated, our testable requirement does not help to determine which specific DMUs do not behave in a cost minimizing manner.

A main novelty of the current paper is that we integrate aspects of efficiency measurement in the nonparametric characterization of cost minimizing production behavior set out above. We include a cost efficiency measure that allows for better discriminating between the performance of different DMUs. We introduce an observation-specific measure of cost efficiency, which we denote by θ_t for every DMU t. The measure captures the deviation between the actually observed input cost at t and the minimal cost for producing the same output, for the given level of unobserved input. Measuring cost efficiency only in terms of observed input costs obtains restrictions that are linear in unknowns, allowing us to make use of standard linear programming techniques to calculate DMU-specific cost efficiencies. Moreover, for our empirical setting, measuring cost

[14]The input bundle $(\mathbf{X_s}, \Omega_s)$ allows production of the output level Q_s. Consider CRS and choose k$=\left(\frac{Q_t}{Q_s}\right)$ in Axiom 1 to obtain that $F(\mathbf{X_s}\left(\frac{Q_t}{Q_s}\right), \Omega_s) = \left(\frac{Q_t}{Q_s}\right)F(\mathbf{X_s}, \Omega_s) = Q_t$. See also Theorem 5 in Varian (1984) for a nonparametric characterization of cost minimizing production behavior under CRS.

efficiency exclusively in terms of observed inputs $\mathbf{X}$ allows for an easier interpretation by Infrabel, as it are these inputs that can be most easily adjusted by the management.

Specifically, we use the following modification of our testable requirement in Proposition 1 (using $\theta_t \in \mathbb{R}^+$ for every t):

$$\theta_t \mathbf{W_t X_t} + \Omega_t \leq \mathbf{W_t} \left[\left(\frac{Q_t}{Q_s} \right) \mathbf{X_s} \right] + \Omega_s.$$

Obviously, the control room dataset S will be consistent with cost minimization (as characterized in Proposition 1) only if it satisfies this inequality when using $\theta_t = 1$ for each DMU t. Choosing a value for θ_t strictly below 1 reduces the observed cost level on the left-hand side of the inequality, so weakening the efficiency condition. More generally, lower values of θ_t allow for greater cost inefficiency.

To evaluate the cost efficiency of a given dataset S, we determine the minimally required cost inefficiency (i.e. waste) to rationalize the observed behavior as cost minimizing. We operationalize this idea through a linear program (LP) with as objective function a weighted sum of the DMU-specific efficiencies θ_t, with weights equal to the DMUs observed costs.[15] Particularly, we solve the following LP problem (LP-OP):

$$\max_{\theta_{t|t\in T},\, \Omega_{t|t\in T}} \quad \sum_{t=1}^{T} \theta_t \mathbf{W_t X_t}$$

subject to

$$\forall t, s \in T : \theta_t \mathbf{W_t X_t} + \Omega_t \leqslant \mathbf{W_t} \left[\left(\tfrac{Q_t}{Q_s} \right) \mathbf{X_s} \right] + \Omega_s.$$

The focus on aggregate cost efficiency, constructed as a weighted sum of the DMU-specific cost efficiencies, seems a natural choice in our empirical application, in which every dataset S corresponds to a separate control room and each DMU t represents an hourly production observation. Thus, LP-OP essentially computes the "overall" cost efficiency of the control room that is under evaluation.[16] Appendix A.1 provides numerical examples illustrating the working of the linear program OP-LP to compute DMU-specific cost efficiencies θ_t.

[15] See, for example, Färe and Zelenyuk (2003); Simar and Zelenyuk (2007); Kuosmanen, Cherchye, and Sipiläinen (2006) on measuring the aggregate cost efficiency of a group of DMUs. These authors make a case for using cost shares as DMU-weights in the aggregate cost efficiency measure when the same input prices apply to all DMUs under evaluation, which is effectively the case in our application setting (focusing on a separate dataset S for each different control room). As a sensitivity check, we also considered using equal weighting of DMUs in our empirical application. This showed robustness of our main findings (results available upon request).

[16] Note that our method computes cost efficiency simultaneously for all DMUs instead of iteratively computing an efficiency score for each DMU individually (as is typically done in DEA). This is because the inclusion of unobserved inputs in a standard DEA framework would hinder the construction of a reference technology against which the performance of the different observations is evaluated.

As a final comment, note that LP-OP not only provides estimates of the DMU-specific efficiencies θ_t but also of the DMU-specific unobserved inputs Ω_t. Our following empirical application will use these estimates Ω_t to analyse the importance of unobserved inputs in the production processes of control rooms on different hours of the day. Clearly the values of Ω_t supporting some given efficiency values θ_t need not be uniquely defined since there may be a set of possible specifications of Ω_t that yield the same DMU-specific efficiency values. We refer to Saelens (2021) for a detailed discussion on characterizing this set by computing observation-specific lower and upper bounds on the unobserved input levels Ω_t. For compactness, we will not consider such set identification in our following empirical application.

3.3 Extensions

We conclude this section by presenting two extensions of our basic methodology that will be instrumental in our following empirical application. We begin by showing how to put additional structure on Ω_t in order to improve the identification of θ_t. Second, we explain how we can account for exogenous environmental variables in the empirical analysis.

Adding structure to Ω. Adding structure to Ω constitutes a productive strategy to strengthen the identification of the DMU-specific efficiencies θ_t. As motivated in Section 2, we can reasonably assume a strong correlation between Ω and the use of ARS for subgroups of our dataset. This allows us to impose ordinal structure on the one-dimensional continuous variable Ω for given subsets of the sample, by using ordinal mappings related to ARS usage. By doing so, we can introduce extra structure on the unobserved inputs in the form of additional constraints that define feasible ranges for the unknown Ω_t.

More specifically, we impose the restriction that, within a given subset $\hat{T} \subseteq T$ (representing a week of data for the evaluated control room), Ω is a strictly monotonically increasing function of a bounding variable, denoted by B (representing the use of ARS): for all observations $t, s \in \hat{T}$, if B_t exceeds B_s then it follows that Ω_t must exceed Ω_s. This restriction can be readily included in the analysis by adding the following constraint to LP-OP:

$$\forall t, s \in \hat{T} : \text{If } B_t > B_s \Rightarrow \Omega_t > \Omega_s.$$

Remark that the analyst does not observe the unobserved input Ω. The dataset merely reports information on ARS usage, of which we believe that it relates to the unobserved input value, as explained above. We then aim to use this information to strengthen identification of Ω.

The proposed mapping strategy parallels similar strategies used by Olley and Pakes (1996) and Ackerberg et al. (2015) to identify heterogeneity in unobserved productivity. These authors assumed a one-to-one mapping between productivity and, respectively, investment and materials. Of course, imposing strict monotonicity is just one specific alternative to implement additional structure on Ω. For example, following the idea of Banker and Morey (1986) one could maintain that only DMUs using a higher level of automation (as measured by the bounding variable) can serve as potential comparison partners. Further, future research could study whether approaches relating to the introduction of ordinal variables in DEA (see Cook et al. (1993, 1996); Cook (2011); Kim et al. (1999); Cooper et al. (1999)) may be compatible with our suggested approach. This could be of particular interest when working with an ordinal bounding variable (as opposed to a continuous one in the current setting). More generally, from a conceptual viewpoint imposing additional structure on the unknown Ω is also closely similar to the use of shadow price/weight restrictions in DEA (see, for example, Allen et al. (1997); Podinovski (2004)).

Environmental variables and robust efficiency estimation. In the application setting that we study, TCs work can be complicated or facilitated by exogenous heterogeneity in the working environment. Therefore, we will include traffic density, lagged delays, and workload of the safety personnel as exogenous environmental variables. Including exogenous environmental variables (captured by an r-dimensional vector $\mathbf{Z} \in \mathbb{R}^r$) in our framework is possible by supplementing it with the (conditional) robust order-m approach that was introduced by Cazals, Florens, and Simar (2002), Daraio and Simar (2005, 2007) and Bădin, Daraio, and Simar (2010) in a DEA context. Conveniently, incorporating this order-m approach also robustifies our efficiency estimates to the presence of outlier observations. In essence, the approach benchmarks the observed production behavior of a DMU against a sample of $m \geq 1$ comparison partners, drawn from the observed set of data. This method is readily implemented through subsampling. In every iteration, first a sample of m potential comparison partners is drawn for every DMU t (i.e. every observation has a potentially different set of comparison partners). Next, for each iteration, we estimate cost efficiencies and unobserved input usages by solving LP-OP. The robust measures of θ_t and Ω_t are then computed by taking means over all iterations.

As explained in detail by Daraio and Simar (2005, 2007), the conditional order-m approach uses kernel weighting methods for drawing reference partners, whereby DMU's within a similar environment as the DMU under study will be more likely to be drawn as one of the m comparison partners. Furthermore, the computation of conditional efficiency scores allows the exploration of the impact of exogenous factors Z on DMU performance, even when no prior knowledge is available on the direction of influence.

The practical implementation of this approach relies on kernel weighting methods to compute probability weights for the data resampling. We follow Bădin, Daraio, and Simar (2010) and select optimal bandwidths for the kernel estimation through nonparametric estimation of a conditional distribution function $F(Q, \mathbf{X} \mid \mathbf{Z})$, using a Least Squares Cross Validation criterion (see Li and Racine (2007)).[17] We consider a generalized product kernel that allows for both continuous and discrete data. For the continuous

[17] To compute the bandwidths we used the 'np' package of Hayfield and Racine (2008).

data (i.e. output Q and environmental variables $\mathbf{Z}$) we employ Epanechnikov weighting. Due to limited variation we treat the input variables $\mathbf{X}$ as discrete ordered variables with Li and Racine (2007) weighting.

4 Empirical findings

The great level of detail present in our real-world and purpose-built dataset enables an informative analysis of the internal functioning of traffic management control rooms at an unusually disaggregate level. In our estimates, every hourly observation serves as a DMU and we will measure hourly observed cost efficiency for the team of TCs active during the hour. Moreover, we consider each control room separately, which allows us to remain agnostic about potential technological differences between control rooms. This obtains sample sizes of a little over 3500 DMUs per control room.[18] We remark that an intrinsic feature of the order-m method that we use for robust efficiency estimation (implemented through subsampling) is that it allows for so-called super-efficient DMUs, which are characterized by a robustified efficiency value θ_t that exceeds 1.

Throughout our empirical analysis we maintain the following assumptions. First, our empirical analysis limits attention to the morning (6h-14h) and day (14h-22h) shifts and does not consider the night shift (22h-6h).[19] Moreover, we exclude observations pertaining to weekends and public holidays because of diverging staffing and traffic levels as compared to weekdays (Roets and Christiaens (2015) discuss in detail the differences between working week and weekend observations). Second, we restrict the unobserved inputs (by appending linear restrictions to LP-OP) such that for all DMUs the unobserved cost share (UCS) is situated between 10% and 50% of the total cost. These bounds are simply starting values (the Matlab functions used to compute our estimates require that lower and upper bounds are specified for each of the unknown variables). Looking at Figure 5 learns that these bounds are hardly ever binding.

In the following Section 4.1 we first compare several model specifications to emphasize the importance of accounting for unobserved inputs. We also present descriptive statistics of our nonparametric estimates of cost efficiency. Next, Section 4.2 documents the hourly-varying effect of omitted variable bias on cost efficiency.

[18] We cleaned the data according to the following sampling criteria. We deleted occurrences of staff working some minutes beyond the end of the shift. Furthermore, we trimmed the lower and upper percentile of the output variable, the environmental variables and output per TC. Lastly, remark that we consider a smaller sample size for control rooms 5, 7 and 9 due to data irregularities.

[19] Specifically, night hour observations are characterized by a low volume of passenger traffic as well as low levels of automation use. Because of this, one might at least suggest that night shift observations are active on a different part of the same technology space. Moreover, we argue that TCs active during night hours pursue objectives that are fundamentally different from those pursued by their colleagues active during the morning or day shift. As such, these observations are not directly comparable.

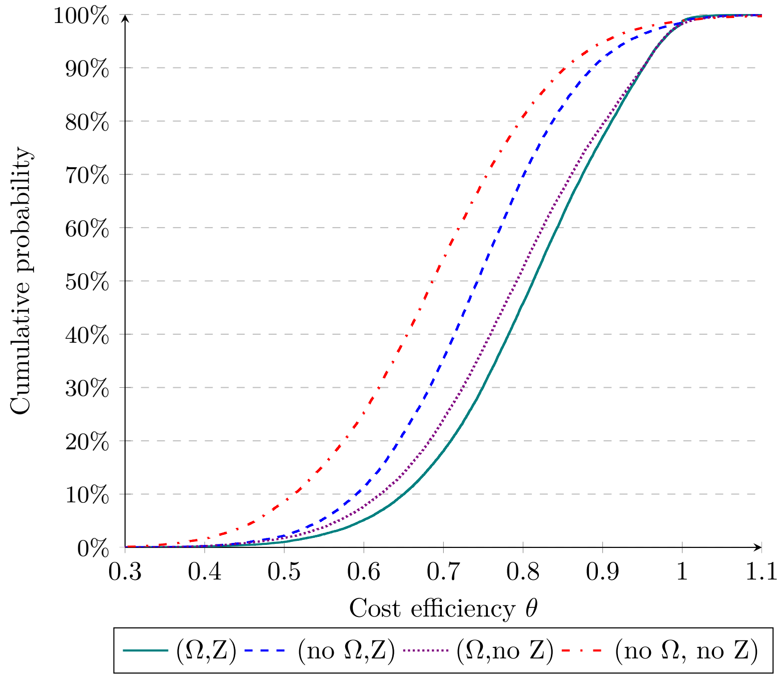
4.1 Comparison between model specifications

We estimated four different specifications of our model. A first version (Specification 1) is similar to existing methods and makes abstraction of both unobserved inputs Ω and environmental variables Z (no Ω, no Z). Here the availability of ARS is neglected and as such the reported results will be subject to an omitted variable bias. Next, in the second version (Specification 2; Ω, no Z) we overcome this problem by including unobserved inputs but not environmental variables. This allows us to filter away the influence of ARS usage on the estimated cost efficiencies. Third, we additionally account for environmental variables (Specification 3; both Ω and Z), isolating the impact of factors exogenous to the production process. The final version documents the impact of accounting for environmental variables Z while making abstraction of unobserved inputs Ω (Specification 4; no Ω, Z). More precisely, for Specifications 1 and 2 (Specifications 3 and 4) we present (conditional) robust estimates computed over 200 iterations. After applying the procedure described in Simar (2003), we chose m=50 to compute our order-m efficiency estimates.

Figure 2 shows the empirical cumulative distribution function of the cost efficiency estimates under our four model specifications, aggregated over the nine control rooms. We observe a substantial difference between the red, dash-dotted line (no Ω, no Z) and the purple, dotted line (Ω, no Z) as well as between the blue, dashed line (no Ω, Z) and the green, solid line (Ω, Z). This evidence suggests that the addition of Ω positively impacts the efficiency scores.[20] Moreover, Kolmogorov-Smirnov tests strongly reject the null hypothesis of identical distributions when Ω is either neglected or accounted for (irrespective of the inclusion of environmental variables Z). Second, the distance between the red, dash-dotted line (no Ω, no Z) and the blue, dashed line (no Ω, Z) as well as between the purple, dotted line (Ω, no Z) and the green, solid line (Ω, Z) shows the impact of controlling for environmental variables Z on efficiency. Clearly, the effect is less pronounced compared to including information on unobserved inputs Ω. Kolmogorov-Smirnov tests indicate that controlling for environmentals leads to significantly larger efficiencies (irrespective of whether unobserved inputs Ω are accounted for). Together, these findings suggest that, for our setting of focus, controlling for unobserved input usage is more informative in explaining variation of efficiency scores than taking into account environmental conditions.

[20] While it is obvious that industry cost efficiency will increase when accounting for Ω, this does not necessarily imply rising cost efficiency for each individual DMU. For example, the empirical application learns that including unobserved input Ω causes a drop in cost efficiency for nearly 1 in 5 DMUs (compared to the setting where Ω is not accounted for).

Figure 2: Empirical cdf of cost efficiency in alternative model specifications - all control rooms



Next, Table 4 bundles descriptive statistics of the computed robust cost efficiencies for the different control rooms, under the different model specifications. The first columns in the table show respectively the mean, standard deviation, median and maximum values of the robust cost efficiencies θ_t^m. In every cell, the row number links up with the model version (e.g. row 1 presents results for Specification 1 (no Ω, no Z), etc.). In general, efficiency is lowest under Specification 1 (no Ω, no Z) with control rooms 2, 5 and 7 reporting mean robust cost efficiencies that slightly exceed 0.60. These low scores reflect the model misspecification and omitted variable bias that is at play because ARS use is not controlled for. As such, we obtain artificially large differences in efficiency across observations. Next, controlling for unobserved input usage in Specification 2 (Ω, no Z) drastically increases cost efficiency of all 9 control rooms. The effect of unobserved inputs is most pronounced for control rooms 5 and 7, of which the mean cost efficiencies increase by no less than 13% points. The remaining control rooms witness an increase in mean cost efficiency of around 8-11% points. In Section 4.2 we will discuss in more detail the resulting changes in hourly cost efficiencies and establish that the omitted variable bias is non-constant across hours-of-the-day.

The presented findings are in line with Stigler (1976)'s argument and should not come as a surprise. The presence of automation and the ARS absorbs part of the hourly variability in workload and helps to better manage unforeseen events, thereby reducing volatility in efficiency levels. This is also reflected in column 4 of Table 4, showing a decrease in the maximum degree of super-efficiency after controlling for unobservables. Given the safety-critical nature of railway traffic control these super-efficient observations (which are characterized by an unusually high workload) may be of special interest to local management. Nonetheless, sizeable inefficiencies persist even when controlling for unobserved inputs. Furthermore, we see from the results of Specification 3 (with both Ω and Z) that these cannot be attributed to merely environmental conditions in the production process. In fact, accounting for environmental factors increases efficiency by only a few % points after accounting for unobserved input. Taken together, these findings suggest that incorporating unobservables into the analysis is far more important for explaining variation in efficiency levels than are environmental factors (at least in the current setting). It seems strongly advisable to account for both environmental factors and unobserved input. As such, in what follows only results of this specification (Ω, Z) will be discussed further. Finally, the results obtained under Specification 4 (no Ω, Z) are found to lie between those of Specification 1 (no Ω, no Z) and Specification 2 (Ω, no Z), again highlighting that accounting for unobserved inputs captures variation in efficiency that cannot be explained by environmental conditions.

In general, in our preferred specification (accounting for both Z and Ω) we observe that efficiency is highest in control rooms 1 and 6. Control rooms 3, 4, 5, 8 and 9 show intermediate performance and control rooms 2 and 7 are characterized by lower average cost efficiencies. These numbers reveal a significant degree of inefficiency within the control rooms, even when controlling for unobserved input usage and environmental factors. Indeed, column 5 of Table 4 shows that only a small fraction of observations has less than 5% potential for observed cost reduction. This need not be a surprise given our discussion in Section 2 mentioning the required margin for unforeseen events and the constraints of the scheduling system as possible sources of inefficiency. In Section 4.2 we will further investigate the possible link between the degree of cost efficiency and hour-of-the-day, which may indicate limited work shift flexibility and suggest to shorten the duration of a shift. Further, the documented variation in efficiency across control rooms may reflect possible idiosyncracies of the work environment. Finally, the last column of Table 4 reports the median value of the standard deviation of the hourly cost efficiencies over the 200 iterations. This shows the extent to which the cost efficiencies vary from one iteration to another, due to variation in the drawing of comparison partners. Accounting for unobserved inputs and, additionally, environmental factors provides a more accurate representation of the production process and as such reduces variability in the estimated efficiency levels.

Table 4: Descriptive statistics of cost efficiencies, by control room

	Version	Mean θ_t^m	Std θ_t^m	Median θ_t^m	Max θ_t^m	# θ_t^m >0.95	Median sd of θ_t
	(no Ω, no Z)	.7633	.1024	.7630	1.1397	145	.0381
Control room 1	(Ω, no Z)	.8338	.0941	.8349	1.1267	505	.0275
(3775 obs)	(Ω, Z)	.8545	.0831	.8573	1.0556	558	.0184
	(no Ω, Z)	.7924	.0915	.7934	1.1617	181	.0334
	(no Ω, no Z)	.6248	.1479	.6147	1.2761	91	.0588
Control room 2	(Ω, no Z)	.7375	.1482	.7356	1.1817	261	.0270
(3548 obs)	(Ω, Z)	.7442	.1467	.7459	1.1807	286	.0257
	(no Ω, Z)	.6656	.1378	.6553	1.2442	115	.0521
	(no Ω, no Z)	.6883	.1192	.6831	1.4781	74	.0769
Control room 3	(Ω, no Z)	.7829	.1207	.7838	1.2209	344	.0305
(3796 obs)	(Ω, Z)	.8045	.1074	.8113	1.1136	292	.0242
	(no Ω, Z)	.7624	.1061	.7651	1.2019	134	.0436
	(no Ω, no Z)	.6645	.1425	.6726	1.3722	81	.0698
Control room 4	(Ω, no Z)	.7744	.1301	.7778	1.1820	343	.0248
(3829 obs)	(Ω, Z)	.7990	.1157	.8040	1.0935	81	.0227
	(no Ω, Z)	.7338	.1126	.7304	1.1340	139	.0462
	(no Ω, no Z)	.6398	.1250	.6281	1.4117	40	.0927
Control room 5	(Ω, no Z)	.7734	.1266	.7714	1.1758	245	.0200
(2587 obs)	(Ω, Z)	.7953	.1077	.7932	1.0566	206	.0264
	(no Ω, Z)	.7462	.1040	.7382	1.2071	93	.0428
	(no Ω, no Z)	.7385	.1054	.7312	1.4157	105	.0526
Control room 6	(Ω, no Z)	.8178	.1050	.8174	1.2014	507	.0142
(3779 obs)	(Ω, Z)	.8308	.0955	.8318	1.1156	482	.0197
	(no Ω, Z)	.7726	.0954	.7677	1.1618	160	.0405
	(no Ω, no Z)	.6251	.1527	.6240	1.3660	39	.0573
Control room 7	(Ω, no Z)	.7550	.1338	.7531	1.1598	110	.0396
(1497 obs)	(Ω, Z)	.7868	.1252	.7836	1.1456	169	.0198
	(no Ω, Z)	.7064	.1317	.7014	1.1768	62	.0359
	(no Ω, no Z)	.6874	.1339	.6856	1.5320	80	.0710
Control room 8	(Ω, no Z)	.7864	.1279	.7905	1.2673	385	.0261
(3786 obs)	(Ω, Z)	.7984	.1199	.8037	1.2840	91	.0225
	(no Ω, Z)	.7283	.1151	.7258	1.2535	111	.0463
	(no Ω, no Z)	.6930	.1207	.6849	1.2125	55	.0554
Control room 9	(Ω, no Z)	.7980	.1122	.7982	1.1841	216	.0286
(2206 obs)	(Ω, Z)	.8146	.1018	.8181	1.2009	187	.0241
	(no Ω, Z)	.7341	.1115	.7260	1.1291	84	.0425

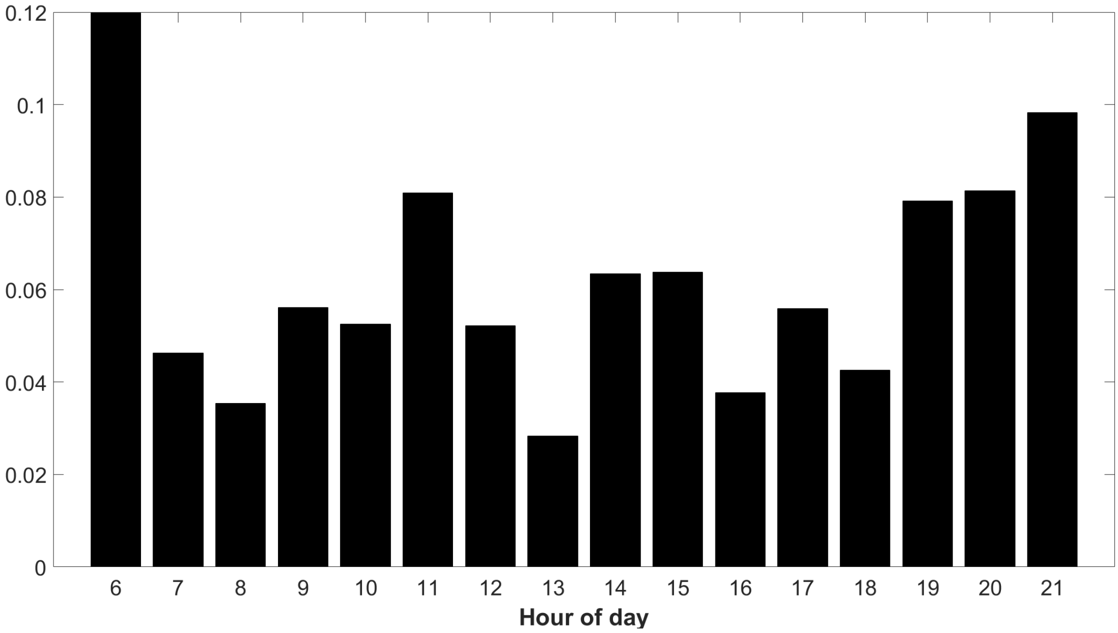
4.2 Highlighting the effects of omitted variable bias

The findings presented in Table 4 indicate a steep increase in average cost efficiency when controlling for unobserved inputs. We now further examine the extent to which the inclusion or exclusion of unobserved input costs causes variation in the estimated cost efficiencies θ and whether this effect may be hour-dependent. To single out the effect of unobserved inputs we compare efficiency across models that both account for environmental differences. Thus, we contrast efficiency scores obtained under the preferred Specification 3 (with both Ω and Z) with those obtained under Specification 4 (no Ω, Z). To keep our discussion focused, the remainder of the paper limits attention to control room 1. Results for the other control rooms can be found in the Appendix A.3.

Because Specification 4 (no Ω, Z) is a very specific case of our preferred Specification 3 (Ω, Z) we expect increases in efficiency when allowing for unobserved inputs. The results in Table 4 confirm this intuition: for control room 1 average robust cost efficiency levels increase by about 6%points when accounting for unobserved input usage. Moreover, Specification 3 (Ω, Z) produces more precise estimates, as can be seen in the last column. Furthermore, this 6% point difference in average robust cost efficiencies conceals significant variation across hours. This is shown in Figure 3 where we subtract the mean hourly robust cost efficiencies θ_t^m obtained under Specification 3 from those obtained in Specification 4, allowing a closer look at the omitted variable bias. As can be seen from Figure 3 the omitted variable bias varies significantly with the hour-of-the-day. More specifically, controlling for unobservables sharply increases efficiency at 6h, 11h and 19h-21h. Interestingly, Figure 1 presented in the introduction learns us that these are hours that are generally characterized by lower output and automation levels. This intuition is further confirmed through the negative correlation of -0.56 between signal automation and the difference in cost efficiency (i.e. the omitted variable bias).

Taken together, our findings suggest that accounting for unobserved inputs leads to a larger increase in efficiency for hours with lower levels of output and automation. This is a clear consequence of the omitted variable bias present in Specification 4 (no Ω, Z). The intuition goes as follows. If we do not account for unobserved input (as in Specification 4), hours with lower output levels will be deemed very inefficient. Afterwards, adding unobservables takes into account that those lower levels of output are mainly produced manually by the TC, while the higher output levels of the other hours are driven to a certain extent by higher levels of automation. As such, in Specification 3 (Ω, Z) we observe a sharp increase in efficiency for those hours with lower output levels.

Figure 3: Control room 1 – Difference between mean hourly θ_t^m obtained under Specification 3 (Ω, Z) and Specification 4 (no Ω, Z)



5 Managerial implications

We conclude our application by illustrating some of the managerial implications of our newly developed methodology. Specifically, we consider two different exercises. First, we study the benefits of educating cross-trained operators (i.e. TCs that are trained to operate in multiple control rooms) to prevent over-performance and super-efficiency. Subsequently, we study our empirical results in more detail and relate variation in cost efficiencies to hourly effects. This will demonstrate a clear need for alternative shift flexibility. While our main focus lies on the interpretation of cost efficiencies, we will also discuss variation in unobserved cost shares.

5.1 Policy implications: cross-trained operators to reduce over-performance

Previously, we mentioned that super-efficient observations may be of particular interest to local management. Given the safety-critical nature of railway traffic control, Infrabel internally cultivates a culture of preventing overly high workload in order to proactively mitigate the possible risks of over-performance. Requiring TCs to work at full mental capacity for multiple consecutive hours presents both health and safety risks (with potential disastrous consequences). For example, previous research established a clear link between TC fatigue and the occurrence of human errors (Roets and Christiaens, 2019; Roets and Folkard, 2022).

The data in column 5 of Table 4 indicate a non-negligible amount of observations obtaining efficiency scores that exceed the 95% level. As such, a significant number of observations may be forced to over-perform. Observe further from Table 4 that the number of nearly efficient units varies considerably depending on whether unobserved inputs are accounted for. Because the omitted variable bias produces larger inefficiency, standard performance measures will tend to systematically underestimate the risk of over-performance. Thus, including unobserved inputs Ω in the analysis is preferable from both a methodological and a safety perspective.

To illustrate the benefits of TCs that are trained to operate in multiple control rooms (i.e. cross-trained operators) we now aim to rebalance the current workforce in a way that alleviates as much as possible the occurrence of super-efficiency and over-performance. Specifically, we propose to reallocate staff from highly inefficient DMUs in the sending control rooms to DMUs in the receiving control room that can be considered prone to over-performing. This while ensuring that the transfer does not make any of the DMUs in the sending control rooms prone to over-performance.[21] Moreover, we impose this rule at the date-hour level so that only DMUs pertaining to the same moment in time can supply staff. This greatly improves realism of the exercise. Conceptually (and technically), this assumes that workers can operate remotely within any control room regardless of their location. In fact, at Infrabel continuous efforts are made to implement such a flexible IT system and to provide staff with the necessary cross-training.

As before, we only discuss results for control room 1. In our preferred Specification 3 (Ω, Z) 260 observations of control room 1 obtain an efficiency level exceeding 0.975, which we consider the threshold for over-performance. Searching for observations with efficiency scores below 0.8 and allowing that merely a single staff member is transferred from any of the sending control rooms to control room 1 remedies over 92% of the over-performance of control room 1 (after controlling that the DMUs in the sending control rooms do not become prone to over-performance). If we additionally allow that only control rooms from the same region (i.e. Flanders or Wallonia) as control room 1 can supply workers (for example because of language differences between regions) then over 81% of over-performance of control room 1 can be avoided. Together, these findings suggest that cross-training and cross-utilization of workers may substantially benefit control rooms.

Finally, we conduct an additional exercise reducing over-performance simultaneously in all nine control rooms. The results indicate that a reorganization of staff (for example because control rooms are made remotely accessible) eliminates 76% of total over-performance across all control rooms. Allowing that only control rooms from within the

[21] Following the transfer we compute counterfactual efficiency scores for the impacted DMUs by comparing the efficient level of observed input costs against the observed input costs reflecting the staff adjustment.

same region can exchange staff, still removes total over-performance by around 62%.

5.2 Assessment of cost efficiency and unobserved costs

Before discussing the results in more detail we present in Table 5 Pearson correlations of the different variables used in the analysis. We computed the input variable by summing staff members, regardless of their grade. The results in Table 5 are rather intuitive. For example, we find that the robust cost efficiencies θ_t^m are correlated negatively with input but positively with output levels, which makes sense. Moreover, we report a considerable correlation between unobserved input costs Ω and both the signal automation variable and output, reflecting the use of signal automation as a bounding variable to restrict variation in unobserved input costs. The strong correlation between unobserved input costs and output further confirms our intuition of the ARS explaining part of the hourly variation in output levels. Lastly, the observed input levels show almost no correlation with output, which is probably due to input levels being constant within 8-hour shifts.

Table 5: Correlation table, control room 1

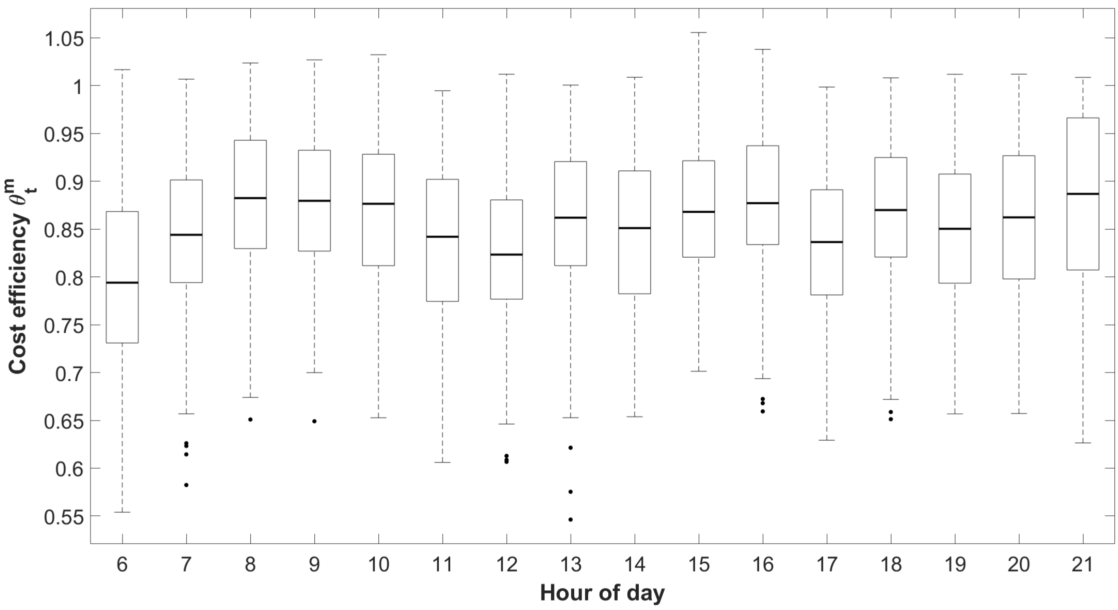
	θ_t^m	Ω	Aggregate input	OutputTC	Signal automation
θ_t^m	1				
Ω	-.2319	1			
Aggregate Input	-.2282	-.2048	1		
OutputTC	.5029	.5627	.0391	1	
Signal automation	-.0809	.7245	.0794	.5439	1

We are now ready to document hourly variation in the reported results, which may provide insights to managers about the sources of inefficiency. Remember that the morning shift runs from 06h in the morning to 14h in the afternoon. The day shift then begins at 14h and runs until 22h. Within shifts staffing levels remain constant, which means we can interpret sudden drops in cost efficiency across hours as signs of limited work shift flexibility (i.e. there being too much staff available for the amount of work that needs to be done).

Hourly variation in cost efficiency. The boxplots presented in Figure 4 indicate how the estimated robust cost efficiencies for control room 1 vary throughout the day. Results for the remaining control rooms are available in Appendix A.3. We use the results for control room 1 to explain the general pattern across control rooms. In general, we establish increasing efficiency levels during the morning rush hours, followed by a common and rather significant drop in efficiency around noon. The reported pattern presents a clear consequence of the existing, inflexible shift structure: although the volume of railway traffic declines during lunch hours, staffing levels still reflect working conditions of the morning rush hours. Afterwards, control rooms exhibit a more ambiguous pattern in the afternoon and late afternoon, showing either a continuous increase or continuous decrease in cost efficiency, with common dips around 16h or 17h. We do not find evidence in favour of increasing efficiency during the late afternoon rush hours, as is the case in the morning shift. Lastly, control room 1 exhibits a rise in efficiency during the final hours of the day. This is likely attributable to trains being directed to centralized parking spots to spend the night, which creates extra workload for the traffic controllers.

Summarizing, the observed pattern of variation in hourly efficiency indicates that revising the starting times (and possibly, the length) of shifts or introducing overlapping shifts may enhance efficiency. For example, to avoid over-presence in control room 1 at 06h and during lunch hours, one possibility may be to have only a proportion of TCs to begin work at 06h, while others end before lunch hours. As such, fewer staff can be present at 06h and during lunch hours. More generally, our method facilitates experimenting with various counterfactual analyses (e.g. overlapping shifts), the outcomes of which may help managers in devising suitable improvement actions. Such exercises, however, fall beyond the scope of the current paper.

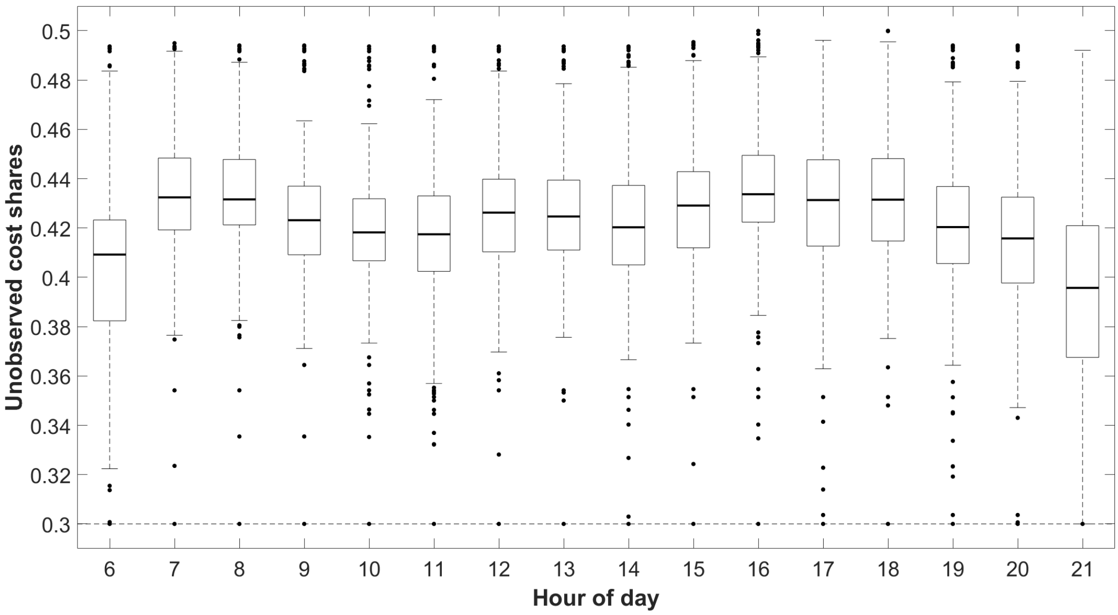
Figure 4: Control room 1 – Boxplots of robust cost efficiencies θ_t^m, by hour



Assessing unobserved costs Ω. Finally, we discuss estimates of the unobserved input costs Ω. In order to obtain a more informative assessment we express the unobserved input costs as a fraction of total costs, which we define as the sum of the observed input costs and Ω. The resulting fractions can be interpreted as unobserved input cost shares (=UCS), naturally bounded between 0 and 1 for every DMU in every iteration. We compute DMU-specific robust UCS by taking an average over the different iterations.

[Figure 5](#) presents boxplots of the robust unobserved cost shares for control room 1, by hour.[22] In general, unobserved input costs represent between 35% and 45% of total costs, showing limited variation across hours, with only a few exceptions. We notice for example a steep increase in (median) UCS moving from 06h to 07h (i.e. the beginning of morning peak hours) and observe a clear spike during afternoon peak hours followed by a drop from 18h onwards. As such, the pattern of hourly variation in the reported UCS in Figure 5 closely follows that of output and automation levels depicted in Figure 1. This result reflects our bounding variable approach which reveals more frequent use of automation during periods of high workload. In this regard, future work could study whether the suggestions proposed above to increase cost efficiency might cause the UCS to converge across hours (especially at 6h an 21h).

Figure 5: Control room 1 – Boxplot of robust unobserved cost shares, by hour



6 Conclusion

Applications of efficiency analysis typically assume that the analyst has full knowledge of all input factors relevant for production. However, one can easily identify production environments where these conditions are hard to meet. As an example, many business operations are characterized by a heavy reliance on intangible input usage, digitization and automation whose intangible nature stands in contradiction to these prior beliefs.

[22] For visual convenience, we only show unobserved cost shares of 30% and above.

As such, existing methods for efficiency analysis and benchmarking may suffer from an omitted variable bias, due to correlation between observed input usage and unobserved (endogenous) inputs.

In the current paper we relax this standard assumption and develop a nonparametric efficiency measurement methodology that is robust to endogeneity issues. We build on the nonparametric cost minimization framework of Cherchye et al. (2021) that uses minimal assumptions to address identification of unobserved inputs. We extend this framework by integrating efficiency measurement tools to evaluate suboptimal conduct. As such, we are the first to differentiate between cost inefficiency (i.e. waste, suboptimal conduct) and unobserved input usage (i.e. optimally chosen input factors that remain unobserved to the analyst). Our approach relates the OR/MS literature on efficiency analysis with the economic literature on heterogeneity in productivity, two literatures which have been regarded as mutually exclusive ever since the influential article of Stigler (1976).

We show the applicability and usefulness of our new approach through an empirical application to customised Belgian railway traffic control data. First, our findings indicate that performance heterogeneity may reflect both inefficient behavior as well as unobserved inputs. Specifically, we report that controlling for unobserved input usage does not preclude the existence of substantial levels of inefficiency. We further show that these resulting inefficiencies cannot be explained by differences in environmental conditions influencing the production process. Second, our results help reveal the channel through which the omitted variable bias affects efficiency and indicate that observations using a lower degree of automation are affected disproportionately if unobserved input usage is neglected. Third, we demonstrate that overlooking the use of unobserved inputs underestimates the risks of over-performance, which may lead to serious concerns from an employee well-being, quality of service and safety perspective. Lastly, we established a clear link between cost efficiency and hour-of-the-day, indicating the potential of cross-training and more customised working arrangements.

The ideas developed in the previous sections serve as a basis for analyzing performance in contemporaneous settings characterized by unobserved input factors. In our opinion, a most interesting extension pertains to our focus on a single-output setting in the present study. The multi-output generalization of our methodology could help to further pinpoint output-specific sources of inefficiency. In this respect, a productive starting point is the nonparametric framework for analyzing multi-output efficiency (with only observed inputs) that is developed in Cherchye, De Rock, Dierynck, Roodhooft, and Sabbe (2013). These authors adopt a similar cost minimization orientation as we do in the current paper.

References

M. A. Abernethy, H. C. Dekker, and J. Grafton. The influence of performance measurement on the processual dynamics of strategic change. *Management Science*, 67(1): 640–659, 2021.

D. A. Ackerberg, K. Caves, and G. Frazer. Identification properties of recent production function estimators. *Econometrica*, 83(6):2411–2451, 2015.

Z. Aksin, M. Armony, and V. Mehrotra. The modern call center: A multi-disciplinary perspective on operations management research. *Production and Operations Management*, 16(6):665–688, 2007.

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

L. Bădin, C. Daraio, and L. Simar. Optimal bandwidth selection for conditional efficiency measures: A data-driven approach. *European Journal of Operational Research*, 201 (2):633–640, 2010.

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

R. D. Banker and A. Maindiratta. Nonparametric analysis of technical and allocative efficiencies in production. *Econometrica: Journal of the Econometric Society*, pages 1315–1332, 1988.

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

N. Bloom and J. Van Reenen. Measuring and explaining management practices across firms and countries. *The Quarterly Journal of Economics*, 122(4):1351–1408, 2007.

E. Brynjolfsson and K. McElheran. The rapid adoption of data-driven decision-making. *American Economic Review*, 106(5):133–39, 2016.

E. Brynjolfsson, D. Rock, and C. Syverson. The productivity J-curve: How intangibles complement general purpose technologies. *American Economic Journal: Macroeconomics*, 13(1):333–372, 2021.

G. M. Campbell. Cross-utilization of workers whose capabilities differ. *Management Science*, 45(5):722–732, 1999.

C. Cazals, J.-P. Florens, and L. Simar. Nonparametric frontier estimation: a robust approach. *Journal of Econometrics*, 106(1):1–25, 2002.

C. Cazals, F. Fève, J.-P. Florens, and L. Simar. Nonparametric instrumental variables estimation for efficiency frontier. *Journal of Econometrics*, 190(2):349–359, 2016.

A. Charnes, W. W. Cooper, and E. Rhodes. Measuring the efficiency of decision making units. *European Journal of Operational Research*, 2(6):429–444, 1978.

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

L. Cherchye, T. Demuynck, B. De Rock, C. Duprez, G. Magerman, and M. Verschelde. Structural identification of productivity under biased technological change. *KU Leuven, Department of Economics Discussion Paper Series*, 2021.

W. D. Cook. Qualitative data in dea. In *Handbook on data envelopment analysis*, pages 151–172. Springer, 2011.

W. D. Cook, M. Kress, and L. M. Seiford. On the use of ordinal data in data envelopment analysis. *Journal of the Operational Research Society*, 44(2):133–140, 1993.

W. D. Cook, M. Kress, and L. M. Seiford. Data envelopment analysis in the presence of both quantitative and qualitative factors. *Journal of the operational research society*, 47(7):945–953, 1996.

W. W. Cooper, K. S. Park, and G. Yu. Idea and ar-idea: Models for dealing with imprecise data in dea. *Management science*, 45(4):597–607, 1999.

J. M. Cordero, D. Santín, and G. Sicilia. Testing the accuracy of dea estimates under endogeneity through a monte carlo simulation. *European Journal of Operational Research*, 244(2):511–518, 2015.

C. Daraio and L. Simar. Introducing environmental variables in nonparametric frontier models: a probabilistic approach. *Journal of Productivity Analysis*, 24(1):93–121, 2005.

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

D. De Prins, L. Simar, and H. Tulkens. Measuring labour efficiency. *Post Offices in The Performance of Public Enterprises: Concepts and Measurement (P. Pestieau ve H. Tulkens M. Marchand)*, pages 243–267, 1984.

U. Doraszelski and J. Jaumandreu. Measuring the bias of technological change. *Journal of Political Economy*, 126(3):1027–1084, 2018.

R. Färe and V. Zelenyuk. On aggregate farrell efficiencies. *European Journal of Operational Research*, 146(3):615–620, 2003.

B.-H. Gong and R. C. Sickles. Finite sample evidence on the performance of stochastic frontiers and data envelopment analysis using panel data. *Journal of Econometrics*, 51(1-2):259–284, 1992.

T. Hayfield and J. S. Racine. Nonparametric econometrics: The np package. *Journal of Statistical Software*, 27(5):1–32, 2008.

J. R. Hicks. *Value and Capital.* 2nd edition, Oxford: The Clarendon Press, 1946.

S.-H. Kim, C.-G. Park, and K.-S. Park. An application of data envelopment analysis in telephone officesevaluation with partial data. *Computers & Operations Research*, 26 (1):59–72, 1999.

T. Kuosmanen, L. Cherchye, and T. Sipiläinen. The law of one price in data envelopment analysis: Restricting weight flexibility across firms. *European Journal of Operational Research*, 170(3):735–757, 2006.

H. Leibenstein. Allocative efficiency vs." x-efficiency". *The American Economic Review*, 56(3):392–415, 1966.

Q. Li and J. S. Racine. *Nonparametric econometrics: theory and practice.* Princeton University Press, 2007.

J. Marschak and W. H. Andrews. Random simultaneous equations and the theory of production. *Econometrica, Journal of the Econometric Society*, pages 143–205, 1944.

F. Nagle. Open source software and firm productivity. *Management Science*, 65(3): 1191–1215, 2019.

S. G. Olley and A. Pakes. The dynamics of productivity in the telecommunications equipment industry. *Econometrica*, 64(6):1263–1297, 1996.

L. Pachl. *Railway operation and control.* Mountlake Terrace, WA: VTD Rail Publishing, 2009.

V. Podinovski. Production trade-offs and weight restrictions in data envelopment analysis. *Journal of the Operational Research Society*, 55(12):1311–1322, 2004.

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

B. Roets and J. Christiaens. Shift work, fatigue, and human error: An empirical analysis of railway traffic control. *Journal of Transportation Safety & Security*, 11(2):207–224, 2019.

B. Roets and S. Folkard. Estimating hourly work schedule risk in railway traffic controllers. *Safety Science*, 151, 2022.

B. Roets, M. Verschelde, and J. Christiaens. Multi-output efficiency and operational safety: An analysis of railway traffic control centre performance. *European Journal of Operational Research*, 271(1):224–237, 2018.

J. Ruggiero. A comparison of dea and the stochastic frontier model using panel data. *International Transactions in Operational Research*, 14(3):259–266, 2007.

D. Saelens. *Essays on nonparametric consumption and production analysis.* PhD thesis, KULeuven, 2021.

D. Santín and G. Sicilia. Dealing with endogeneity in data envelopment analysis applications. *Expert Systems with Applications*, 68:173–184, 2017.

A. Saunders and E. Brynjolfsson. Valuing information technology related intangible assets. *MIS Quarterly*, 40(1):83–110, 2016.

L. Simar. Detecting outliers in frontier models: A simple approach. *Journal of Productivity Analysis*, 20(3):391–424, 2003.

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

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

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

L. Simar, A. Vanhems, and I. Van Keilegom. Unobserved heterogeneity and endogeneity in nonparametric frontier estimation. *Journal of Econometrics*, 190(2):360–373, 2016.

G. J. Stigler. The xistence of x-efficiency. *The American Economic Review*, 66(1): 213–216, 1976.

C. Syverson. What determines productivity? *Journal of Economic literature*, 49(2): 326–65, 2011.

T. F. Tan and S. Netessine. When does the devil make work? an empirical study of the impact of workload on worker productivity. *Management Science*, 60(6):1574–1593, 2014.

T. G. Topcu, K. Triantis, and B. Roets. Estimation of the workload boundary in socio-technical infrastructure management systems: The case of belgian railroads. *European Journal of Operational Research*, 278(1):314–329, 2019.

H. R. Varian. The nonparametric approach to production analysis. *Econometrica: Journal of the Econometric Society*, pages 579–597, 1984.

L. Wu, L. Hitt, and B. Lou. Data analytics, innovation, and firm productivity. *Management Science*, 66(5):2017–2039, 2020.

A Appendix

A.1 Numerical examples

The following examples illustrate the working of our linear program OP-LP to compute the DMU-specific cost efficiencies θ_t. We show that cost minimizing behavior can be rejected even in minimalistic setting with information on only two DMUs, making use of only 2 observed inputs.

Our first example uses a dataset S with input prices $W_1 = (1,2)$ and $W_2 = (2,1)$ and input quantities $X_1 = (1,2)$ and $X_2 = (2,1)$. For given cost efficiency values θ_1 and θ_2, the restrictions of OP-LP state that

$$5\theta_1 + \Omega_1 \leq \left(\frac{Q_1}{Q_2}\right) * 4 + \Omega_2,$$

$$5\theta_2 + \Omega_2 \leq \left(\frac{Q_2}{Q_1}\right) * 4 + \Omega_1.$$

We can reformulate this as

$$\Omega_1 - \Omega_2 \leq \left(\frac{Q_1}{Q_2}\right) * 4 - 5\theta_1 \text{ and}$$

$$\Omega_1 - \Omega_2 \geq -\left(\frac{Q_2}{Q_1}\right) * 4 + 5\theta_2,$$

which implies that $\theta_1 + \theta_2 \leq \frac{4}{5}\left(\frac{Q_1}{Q_2} + \frac{Q_2}{Q_1}\right)$. If we then assume that the output levels Q_1 and Q_2 are such that

$$\frac{1}{2} < \frac{Q_1}{Q_2} < 2,$$

we obtain that there do not exist positive values for Ω_1 and Ω_2 that make both DMUs cost efficient (i.e. $\theta_1 = \theta_2 = 1$). To take a specific instance, assume that $Q_1 = 30$ and $Q_2 = 40$. Then, solving OP-LP will yield the efficiency values $\theta_1 = 1$ and $\theta_2 = 0.67$ (corresponding to $\Omega_1 = 0$ and $\Omega_2 = 2$).

Our second example assumes the same input prices for all DMUs under evaluation, which also applies to our own empirical application. In this case, we obtain testable implications when information is available on only three DMUs, making use of only 2 observed inputs. To see this, consider a dataset S with input prices $W = (1,2)$ and input quantities $X_1 = (1,2)$, $X_2 = (2,1)$ and $X_3 = (2,2)$. For given cost efficiency values of θ_1, θ_2 and θ_3, the restrictions of OP-LP imply

$$5\theta_1 + \Omega_1 \leq \left(\frac{Q_1}{Q_2}\right) * 4 + \Omega_2,$$

$$5\theta_1 + \Omega_1 \leq \left(\frac{Q_1}{Q_3}\right) * 6 + \Omega_3,$$

$$4\theta_2 + \Omega_2 \leq \left(\frac{Q_2}{Q_1}\right) * 5 + \Omega_1,$$

$$4\theta_2 + \Omega_2 \leq \left(\frac{Q_2}{Q_3}\right) * 6 + \Omega_3,$$

$$6\theta_3 + \Omega_3 \leq \left(\frac{Q_3}{Q_1}\right) * 5 + \Omega_1,$$

$$6\theta_3 + \Omega_3 \leq \left(\frac{Q_3}{Q_2}\right) * 4 + \Omega_2.$$

Adding up these inequalities obtains

$$4\theta_2 + 5\theta_1 \leq \left(\frac{Q_1}{Q_2}\right) * 4 + \left(\frac{Q_2}{Q_1}\right) * 5,$$

$$6\theta_3 + 5\theta_1 \leq \left(\frac{Q_1}{Q_3}\right) * 6 + \left(\frac{Q_3}{Q_1}\right) * 5,$$

$$6\theta_3 + 4\theta_2 \leq \left(\frac{Q_2}{Q_3}\right) * 6 + \left(\frac{Q_3}{Q_2}\right) * 4.$$

For $\left(\frac{Q_2}{Q_3}\right) = \frac{3}{4}$ we obtain inconsistency with the lastly mentioned constraint for $\theta_1 = \theta_2 = \theta_3 = 1$. As such, there do not exist positive values for Ω_1, Ω_2 and Ω_3 that render this dataset S cost efficient. As a specific instance, let $Q_1 = 50$, $Q_2 = 30$ and $Q_3 = 40$. Then, solving OP-LP gives the efficiency values $\theta_1 = 1$ and $\theta_2 = 1$ and $\theta_3 = 0.9722$ (corresponding to $\Omega_1 = 1.8333$, $\Omega_2 = 0.5$ and $\Omega_3 = 0$).

A.2 Monte Carlo simulation

We simulate a production setting that is characterized by unobserved input to analyze the empirical performance of our efficiency evaluation model by benchmarking it against the standard DEA model with constant returns to scale (i.e. the so-called CCR model, after Charnes et al. (1978)).

Set-up. We assume the same data generating process as Cherchye et al. (2021) in their original study. These authors consider a standard set-up with two inputs (capital and labor) characterized by a Constant Elasticity of Substitution production function and characterized by a labor bias (see also Doraszelski and Jaumandreu (2018)):

$$Q = (\Omega \left[\alpha K^\rho + (1-\alpha)(\Omega^\delta L)^\rho\right]^{\frac{v}{\rho}})e^\epsilon. \tag{1}$$

The addition of a technological bias in labor assures that our method is not limited to Hicks-neutral technological change. In this function Q represents output, Ω is an unobserved productivity input. K and L represent observed capital and labor usage, respectively. Previous studies have demonstrated that nonparametric efficiency measures are sensitive to measurement error (see e.g. Gong and Sickles (1992); Ruggiero (2007)). To analyze the robustness of our method to different levels of measurement error we generate an error term ϵ representing mean-zero normally distributed noise with variance σ_ϵ^2. We consider 3 different level of noise with $\sigma_\epsilon^2 = \{0, 0.025, 0.05\}$. In addition, we set $\delta = -0.5$, $\rho = 0.75$, $\alpha = 0.5$ and $v = 1$ (=CRS) to calibrate the model. Ω and K follow a jointly normal distribution. We consider 2 different scenario's for the variance-covariance matrix:

$$\text{Scenario a):} \begin{pmatrix} \Omega \\ K \end{pmatrix} \sim \mathcal{N}_2 \left[\begin{pmatrix} 5 \\ 5 \end{pmatrix}, \begin{pmatrix} 0.1 & -0.01 \\ -0.01 & 1 \end{pmatrix} \right],$$
$$\text{Scenario b):} \begin{pmatrix} \Omega \\ K \end{pmatrix} \sim \mathcal{N}_2 \left[\begin{pmatrix} 5 \\ 5 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \right].$$

In scenario a) there is heterogeneity in unobserved input Ω and we allow for some degree of correlation between capital inputs and the unobserved input (note that there will also be correlation between Ω and the labor input L). As such, there is potential for an omitted variable bias. By contrast, in scenario b), there is no heterogeneity in unobserved input usage and thus no omitted variable bias. Ω is simply a constant that does not impact the optimization problem. We compare our estimates against the CCR model which we predict will only provide accurate estimates in scenario b) (when the unobserved inputs are essentially non-existent).

We set $w_K = 2$ and $w_\Omega = 4$ for all observations. Using these generated values $(\Omega, w_\Omega, K, w_K, \delta, \alpha, \rho, v)$ we solve for optimal labor usages L, labor prices w_L and output levels Q that correspond with cost-minimizing behavior. Further, these output levels Q are then multiplied by e^ϵ to introduce noise. Finally, we simulate 25% of observations to be efficient $\theta = 1$ with the remaining inefficiencies following an i.i.d. half-normal distribution $\theta \sim exp(- \mid \mathcal{N}(0, 0.02) \mid)$ and generate inefficient input levels through the division of K and L by θ. A bounding variable is additionally calculated to reflect the ordering of the unobserved costs within a group i.

Simulation results. We ran computational experiments with varying sample sizes of 500 and 2000 observations considering three different levels of noise. For each of these 6 cases, we generated 200 Monte Carlo samples for which we provide results below. For simplicity, we abstain from using robust methods. We ran LP-OP with an additional bounding variable that restricts the Ω values within every group i. In all samples, each group i contains 50 non-overlapping observations. This ensures group sizes that are always smaller than the ones considered in the empirical application, where groups comprise around 80 observations. Next, in each sample we impose a common lower (upper) bound on the Ω values lying 10% below (above) the lowest (highest) unobserved cost in the sample. Further, we did not allow for super-efficient observations in the simulations and simply set all $\Omega = 5$ when computing scenario b).

Let us then consider our estimates of the cost efficiencies θ and unobserved inputs Ω.

Table 6 documents the average Spearman rank correlations between the true and estimated values, with the upper (lower) panel reporting simulation results under scenario a (b). Consider first scenario a), denoted in the upper panel of Table 6. Column 1 presents

Table 6: Mean Spearman correlations between true and estimated values

		θ	θ_{CCR}	Ω
Scenario a)				
N=500	σ_ε^2=0	.9477	.8515	.9830
	$\sigma_\varepsilon^2 = 0.025$	.8650	.8171	.9215
	$\sigma_\varepsilon^2 = 0.05$	.7448	.7353	.8478
N=2000	σ_ε^2=0	.9498	.8553	.9821
	σ_ε^2=0.025	.8643	.8206	.9143
	σ_ε^2=0.05	.7432	.7361	.8346
Scenario b)				
N=500	σ_ε^2=0	.9999	.9874	-
	σ_ε^2=0.025	.9249	.9139	-
	σ_ε^2=0.05	.8175	.7988	-
N=2000	σ_ε^2=0	1	.9909	-
	σ_ε^2=0.025	.9272	.9224	-
	σ_ε^2=0.05	.8195	.8091	-

mean Spearman rank correlations between the true θ and their estimates obtained by our method, for different degrees of noise. We observe that our produced model estimates correlate strongly with the true cost inefficiencies. Although, not surprisingly, the correlations deteriorate slightly with the addition of increasing levels of noise, our model outperforms the CCR model, for which correlations with the true values are reported in the second column. Next, column 3 shows Spearman rank correlations between true and estimated unobserved input costs. We observe high correlations between the estimated and true unobserved input cost levels, exceeding the correlations reported for θ. Again, the reported correlation drops with increasing levels of noise.

Now consider the results under scenario b), denoted in the lower panel of Table 6. As expected, both CCR and our model obtain accurate estimates.[23] These results underline the usefulness of our model even when the existence of unobserved inputs is questionable.

The previous results illustrate the usefulness of our approach in relation to existing methods. Next, to assess the quality of our simulation results we computed the Mean Absolute Deviation (=MAD) between the true values and model estimates. The MAD

[23]The outperformance of our model in comparison to the CCR model is due to the inclusion of price information for the observed inputs in our model. We restrict our comparison to the CCR model without inclusion of input price information, as this is the most popularly used DEA methodology in the literature.

shows how well the estimates of the different models approximate the true values (i.e. how close do the estimates lie to their true values). Together with the general impressions obtained from the correlation analysis the MAD allows for a more comprehensive assessment of simulation performance. Table 7 documents the Mean Absolute Deviations between the true and estimated values, with the upper (lower) panel reporting simulation results under scenario a (b). Examining first the upper panel, we find the MAD

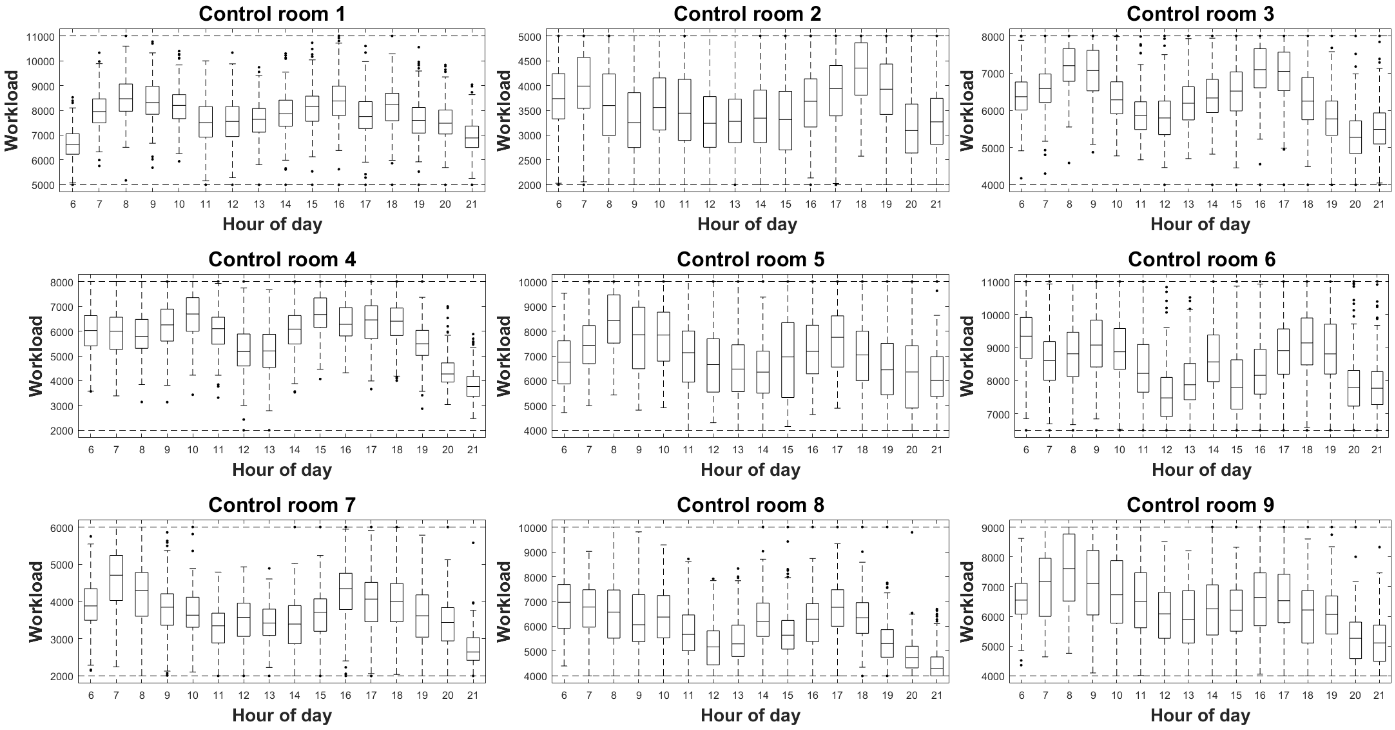
Table 7: Average Mean Absolute Deviation between true and estimated values

		θ (%point)	θ_{CCR} (%point)	Ω (%)
Scenario a)				
N=500	σ_ε^2=0	.0100	.0689	.1244
	$\sigma_\varepsilon^2 = 0.025$	.0228	.0833	.1215
	$\sigma_\varepsilon^2 = 0.05$	.0472	.1165	.1065
N=2000	σ_ε^2=0	.0098	.0854	.1277
	σ_ε^2=0.025	.0231	.1049	.1154
	σ_ε^2=0.05	.0481	.1470	.0916
Scenario b)				
N=500	σ_ε^2=0	4e-5	.0004	-
	σ_ε^2=0.025	.0517	.0461	-
	σ_ε^2=0.05	.1056	.0938	-
N=2000	σ_ε^2=0	8e-6	.0001	-
	σ_ε^2=0.025	.0614	.0568	-
	σ_ε^2=0.05	.1232	.1135	-

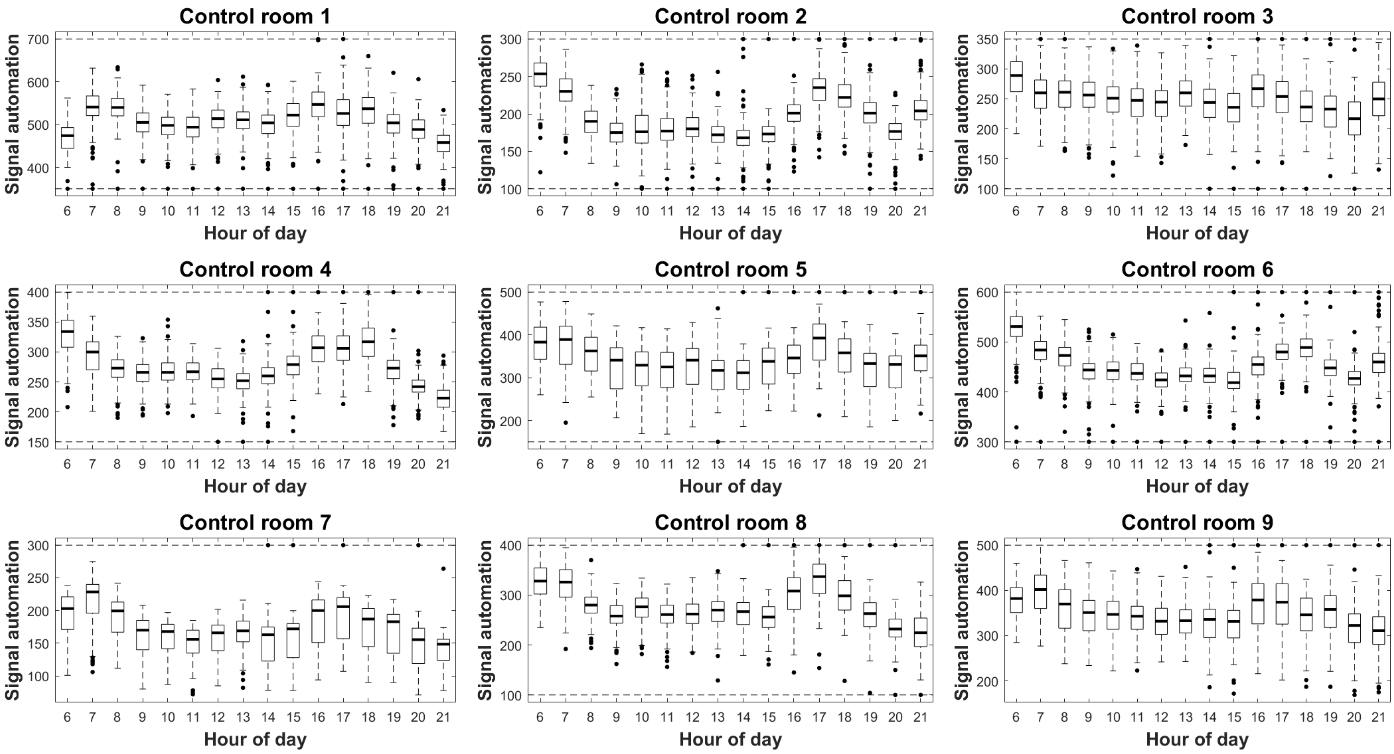
for θ obtained by our method (column 1 in Table 7) to be generally small, although increasing with higher levels of noise. Moreover, these MADs are much smaller than those obtained by the CCR model (presented in column 2). This finding further confirms the previously established outperformance against the CCR model. Next, a comparison of the true unobserved input costs with the estimated values in step 1 (column 3) obtains MADs representing roughly a 10-12% deviation. Finally, under scenario b (lower panel of Table 7), we observe a similar performance for both models in approximating the true θ.

A.3 Additional empirical results

A.3.1 Boxplots of workload and signal automation (all control rooms)



(a) All control rooms: workload, by hour



(b) All control rooms: # of automatically opened signals, by hour

Figure 6: Boxplots of workload and signal automation (all control rooms)

A.3.2 Correlation tables for all control rooms

Table 8: Correlation table, control room 1

	θ_t^m	Ω	Aggregate input	OutputTC	Signal automation
θ_t^m	1				
Ω	-.2319	1			
Aggregate Input	-.2282	-.2048	1		
OutputTC	.5029	.5627	.0391	1	
Signal automation	-.0809	.7245	.0794	.5439	1

Table 9: Correlation table, control room 2

	θ_t^m	Ω	Aggregate input	OutputTC	Signal automation
θ_t^m	1				
Ω	-.2263	1			
Aggregate Input	-.1190	.0279	1		
OutputTC	.6487	.5544	.0490	1	
Signal automation	-0.0260	.6803	.0439	.4764	1

Table 10: Correlation table, control room 3

	θ_t^m	Ω	Aggregate input	OutputTC	Signal automation
θ_t^m	1				
Ω	-.2148	1			
Aggregate Input	-.4697	-.0500	1		
OutputTC	.6243	.4117	-.0549	1	
Signal automation	-.1813	.4898	-.0290	.1574	1

Table 11: Correlation table, control room 4

	θ_t^m	Ω	Aggregate input	OutputTC	Signal automation
θ_t^m	1				
Ω	-.2601	1			
Aggregate Input	-.1369	-.0367	1		
OutputTC	.5098	.6048	.0708	1	
Signal automation	-.0291	.6486	.0554	.5624	1

Table 12: Correlation table, control room 5

	θ_t^m	Ω	Aggregate input	OutputTC	Signal automation
θ_t^m	1				
Ω	-.1670	1			
Aggregate Input	.0687	-.0776	1		
OutputTC	.5791	.4232	.5476	1	
Signal automation	.0829	.5027	.5592	.6993	1

Table 13: Correlation table, control room 6

	θ_t^m	Ω	Aggregate input	OutputTC	Signal automation
θ_t^m	1				
Ω	-.2992	1			
Aggregate Input	-.1665	-.0149	1		
OutputTC	.6174	.4965	.0175	1	
Signal automation	-.0912	.6825	-.0317	.4342	1

Table 14: Correlation table, control room 7

	θ_t^m	Ω	Aggregate input	OutputTC	Signal automation
θ_t^m	1				
Ω	-.0972	1			
Aggregate Input	-.1605	.1649	1		
OutputTC	.5298	.6995	.2543	1	
Signal automation	.0822	.7309	.3461	.6872	1

Table 15: Correlation table, control room 8

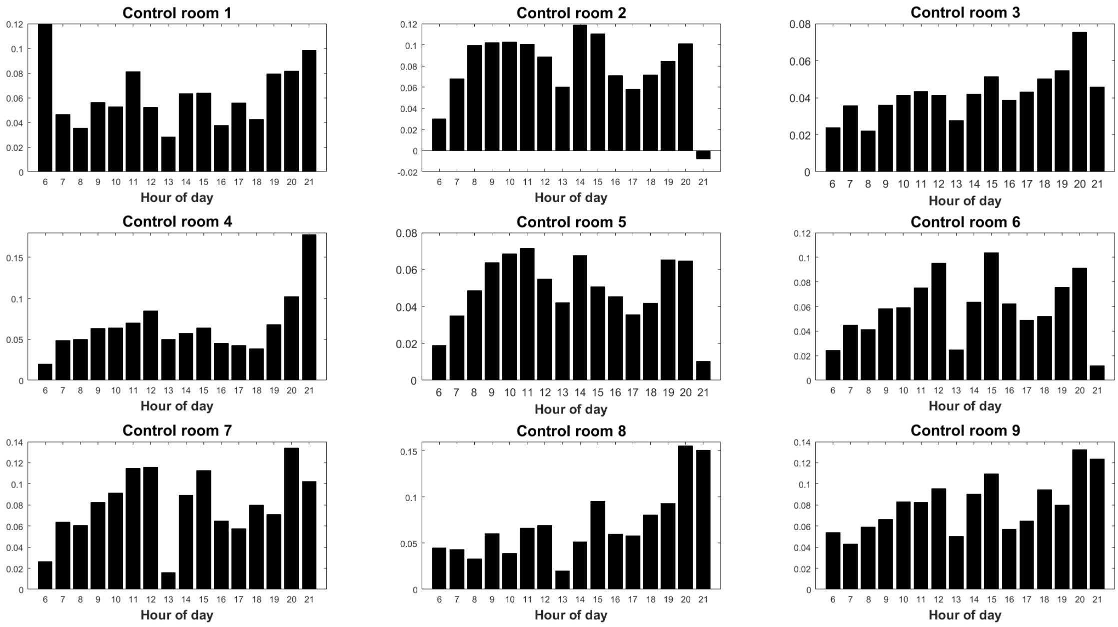
	θ_t^m	Ω	Aggregate input	OutputTC	Signal automation
θ_t^m	1				
Ω	-.2037	1			
Aggregate Input	-.0163	-.2606	1		
OutputTC	.5626	.3840	.4290	1	
Signal automation	.0260	.4785	.4404	.6211	1

Table 16: Correlation table, control room 9

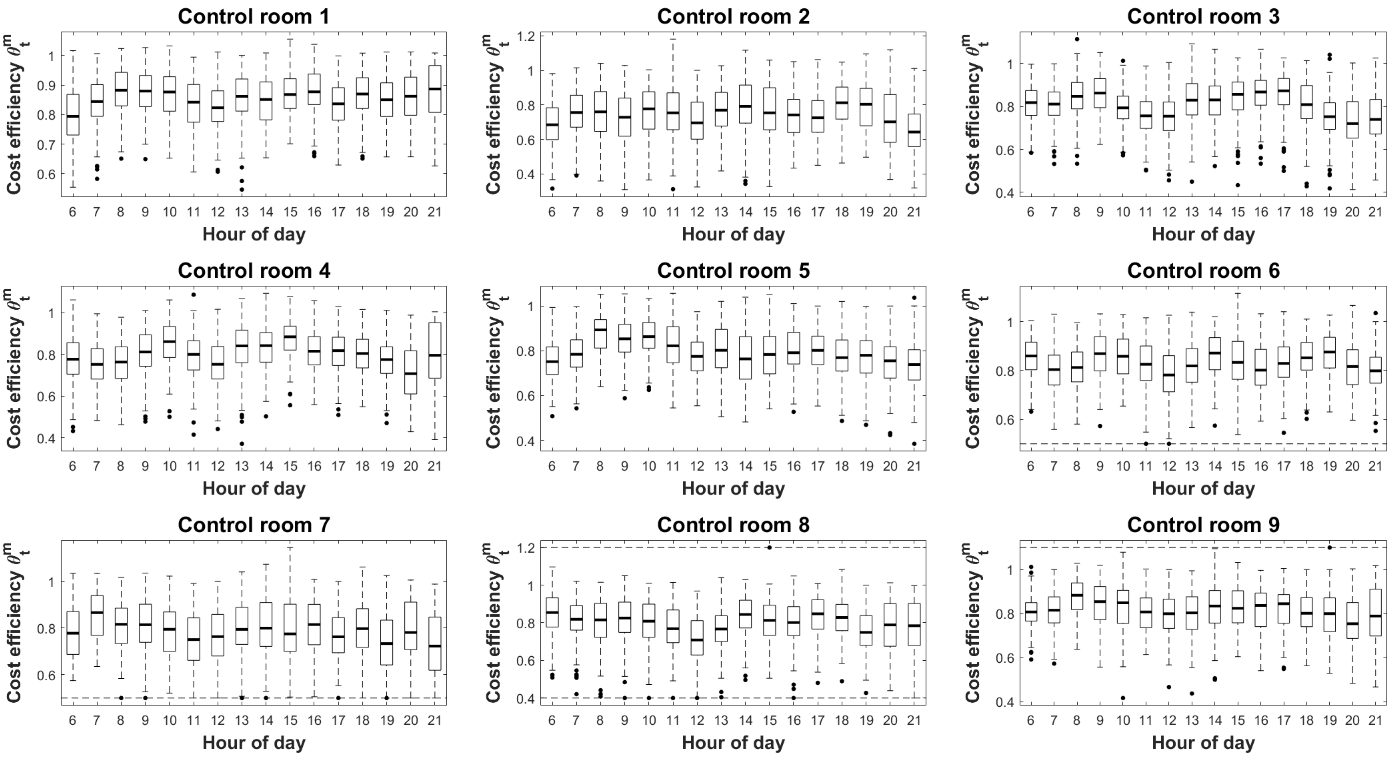
	θ_t^m	Ω	Aggregate input	OutputTC	Signal automation
θ_t^m	1				
Ω	-.2560	1			
Aggregate Input	.1139	-.4331	1		
OutputTC	.5275	.1548	.6077	1	
Signal automation	.1027	.2137	.5930	.7030	1

A.3.3 Highlighting the omitted variable bias (all control rooms)

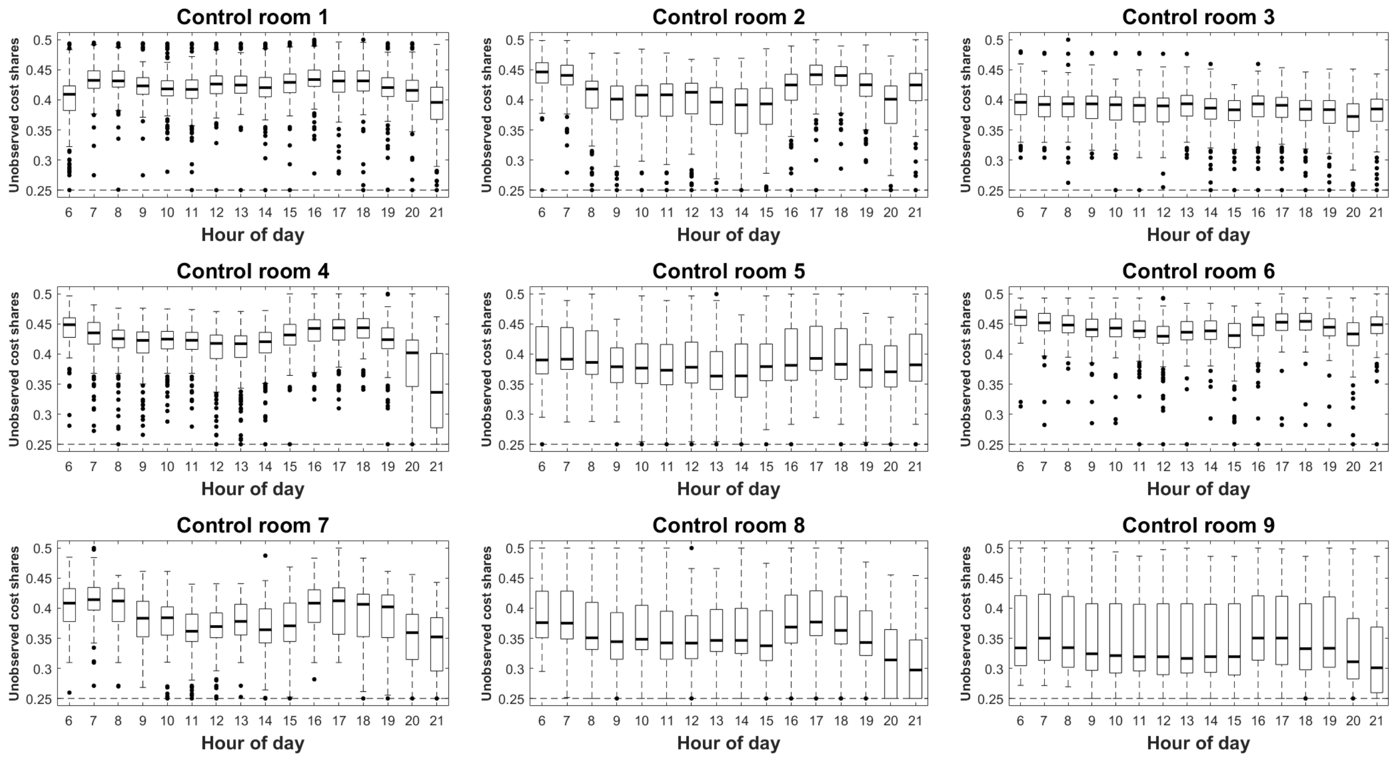
Figure 7: Difference between mean hourly θ_t^m obtained under version 3 and version 4



A.3.4 Assessment of robust cost efficiency and unobserved input costs (all control rooms)



(a) All control rooms: Boxplots of the robust cost efficiencies θ_t^m, by hour



(b) All control rooms: Boxplots of robust unobserved cost shares, by hour