Management of Complex Sociotechnical Systems

Taylan Güneş Topcu

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
In
Industrial and Systems Engineering

Konstantinos Triantis, Chair
Navid Ghaffarzadegan
Alejandro Salado Diez
Paul Collopy
Bart Roets

March 17, 2020
Falls Church, VA

Keywords: sociotechnical systems, infrastructure systems engineering, efficiency performance measurement, data envelopment analysis (DEA), machine learning, autonomous systems.

Management of Complex Sociotechnical Systems

Taylan Güneş Topcu

ABSTRACT

Sociotechnical systems (STSs) rely on the collaboration between humans and autonomous decision-making units to fulfill their objectives. Highly intertwined social and technical contextual factors influence the collaboration between these human and engineered elements, and consequently the performance characteristics of the STS. In the next two decades, the role allocated to STSs in our society will drastically increase. Thus, the effective design of STSs requires an improved understanding of the human-autonomy interdependency.

This dissertation brings together management science along with systems thinking and uses a mixed-methods approach to investigate the interdependencies between people and the autonomous systems they collaborate within complex socio-technical enterprises. The dissertation is organized in three mutually exclusive essays, each investigating a distinct facet of STSs: safe management, collaboration, and efficiency measurement.

The first essay investigates the amount of work allocated to safety-critical decision makers and quantifies Rasmussen's workload boundary that represents the limit of attainable workload. The major contribution of this study is to quantify the qualitative theoretical construct of the workload boundary through a Pareto-Koopmans frontier. This frontier allows one to capture the aggregate impact of the social and technical factors that originate from operational conditions on workload.

The second essay studies how teams of humans and their autonomous partners share work, given their subjective preferences and contextual operational conditions. This study presents a novel integration of machine learning algorithms in an efficiency measurement

framework to understand the influence of contextual factors. The results demonstrate that autonomous units successfully handle relatively simple operational conditions, while complex operational conditions require both workers and their autonomous counterparts to collaborate towards common objectives.

The third essay explores the complementary and contrasting roles of efficiency measurement approaches that deal with the influence of contextual factors and their sensitivity to sample size. The results are organized in a structured taxonomy of their fundamental assumptions, limitations, mathematical structure, sensitivity to sample size, and their practical usefulness.

To summarize, this dissertation provides an interdisciplinary and pragmatic research approach that benefits from the strengths of both theoretical and data-driven empirical approaches. Broader impacts of this dissertation are disseminated among the literatures of systems engineering, operations research, management science, and mechanical design.

Management of Complex Sociotechnical Systems

Taylan Güneş Topcu

GENERAL AUDIENCE ABSTRACT

A system is an integrated set of elements that achieve a purpose or goal. An autonomous system (ADS) is an engineered element that often substitutes for a human decision-maker, such as in the case of an autonomous vehicle. Sociotechnical systems (STSs) are systems that involve the collaboration of a human decision-maker with an ADS to fulfill their objectives. Historically, STSs have been used primarily for handling safety critical tasks, such as management of nuclear power plants. By design, STSs rely heavily on a collaboration between humans and ADS decision-makers. Therefore, the overall characteristics of a STS, such as system safety, performance, or reliability; is fully dependent on human decisions. The problem with that is that people are independent entities, who can be influenced by operational conditions. Unlike their engineered counterparts, people can be cognitively challenged, tired, or distracted, and consequently make mistakes.

The current dependency on human decisions, incentivize business owners and engineers alike to increase the level of automation in engineered systems. This allows them to reduce operational costs, increase performance, and minimize human errors. However, the recent commercial aircraft accidents (e.g., Boeing 737-MAX) have indicated that increasing the level of automation is not always the best strategy. Given that increasing technological capabilities will spread the adoption of STSs, vast majority of existing jobs will either be fully replaced by an ADS or will change from a manual set-up into a STS. Therefore, we need a better understanding of the relationships between social (human) and engineered elements.

This dissertation, brings together management science with systems thinking to investigate the dependencies between people and the autonomous systems they collaborate within

complex socio-technical enterprises. The dissertation is organized in three mutually exclusive essays, each investigating a distinct facet of STSs: safe management, collaboration, and efficiency measurement.

The first essay investigates the amount of work handled by safety-critical decision makers in STSs. Primary contribution of this study is to use an analytic method to quantify the amount of work a person could safely handle within a STSs. This method also allows to capture the aggregate impact of the social and technical factors that originate from operational conditions on workload.

The second essay studies how teams of humans and their autonomous partners share work, given their preferences and operational conditions. This study presents a novel integration of machine learning algorithms to understand operational influences that propel a human-decision maker to handle the work manually or delegate it to ADSs. The results demonstrate that autonomous units successfully handle simple operational conditions. More complex conditions require both workers and their autonomous counterparts to collaborate towards common objectives.

The third essay explores the complementary and contrasting roles of data-driven analytical management approaches that deal with the operational factors and investigates their sensitivity to sample size. The results are organized based on their fundamental assumptions, limitations, mathematical structure, sensitivity to sample size, and their practical usefulness.

To summarize, this dissertation provides an interdisciplinary and pragmatic research approach that benefits from the strengths of both theoretical and data-driven empirical approaches. Broader impacts of this dissertation are disseminated among the literatures of systems engineering, operations research, management science, and mechanical design.

Acknowledgements

Getting a doctorate degree could be tough. However, my journey was an exceptional experience thanks to the support of an amazing group of people for whom I will always be grateful. Below, I would like to take the opportunity to express my appreciation to some of these individuals.

Selin, my soulmate, your exceptional character inspires me every day to do better. Your unconditional love and support is my rock and clearly, none of this would have been possible without you.

Prof. Kostas Triantis, my academic father, thank you for making this a pleasant experience. I will always cherish our conversations, your wisdom, friendship, and the fact that you respected and cared about my interests as a graduate student. I will carry-on your "art" for as long as I live and I hope to shape the minds of others like you shaped mine.

Dr. Bart Roets, I was extremely fortunate to have crossed paths with you. This research was enabled by your excellent communication skills, curiosity, and invaluable expertise. Thank you for trusting me and making yourself available despite the time gap and your other obligations.

Prof. Paul Collopy, thank you for convincing me to pursue a doctorate degree and for everything you have done for me over the years. Your continued mentorship fuels my intellectual curiosity and motivates me to explore new dimensions of my research. I cannot possibly thank you enough for your time and wisdom.

Prof. Alejandro Salado, thank you for everything you taught me and making yourself available whenever I needed guidance. I specifically appreciate your creativity and for training me as your assistant for the fundamentals course.

Prof. Navid Ghaffarzadegan, thank you for your wisdom and kindness. I learned a lot from you regarding the fundamentals of social systems and their behavior over time, which I will always be grateful.

Prof. Van Aken, thank you for your leadership and nurturing this positive learning environment that I am proud to call home. Your continued support, especially during the last turbulent period, will always be remembered.

Prof. Bryan Mesmer, thank you for being an excellent mentor. Thank you for caring and making yourself available to me whenever I needed help.

Dr. Oscar Herrera-Restrepo, my academic elder brother, your mentoring and friendship were comforting even through my darkest graduate school moments.

Prof. Alexandra Medina-Borja, your wisdom, guidance, and positivity has always been a breeze of fresh air.

My dear colleagues and fellow Hokies, *Dr. Marcos Santos, Ning-Yuan "Georgia" Liu, Dr. Alireza Ebrahimvandi, Negar Darabi, Sarah Mostafavi, Glen Lyddane, Jaber Valinejad, Dr. Zhao Junbo, Dr. Yijun Xu, Dr. Roma Bhatkoti, Hesam Mahmoudi, Saman Mohsenirad and Mohammed Ba-Aoum* it was a privilege to get to know you. Thank you for your friendship and for sharing this journey with me.

Hannah Parks, James Murphy, Jeny Beausoliel, and Jessica Mullins you made me feel on campus through your continuous support, thank you!

Before I conclude, *it takes a village to raise a child*. I was extremely fortunate to have a dedicated and loving family that shaped my intellectual development and enabled me to pursue my dreams. My family interactions significantly contributed to my critical thinking abilities and led me to develop an interest in complex systems, which eventually brought me to where I am today. To elaborate:

My mother, who was an activist educator, convinced me that I can accomplish whatever I set my mind onto through discipline, science, and education. She also got me interested in travelling, which in return exposed me to how socioeconomic conditions and cultural biases affect the lives of many.

My father was a civil engineer and he would take me to construction zones ever since I could walk (no he did not think it was too soon). An oddly risky way to entertain your only child, yet I enjoyed observing how structures are built and asking questions about the process. I would also tinker around and play with the work machines, which must have established the foundations of my passion for engineered systems.

My grandmother taught me how to play multiplayer card games when I was five and later allowed me to accompany her during in-house gambling sessions. I learned an awful lot about strategic decision making under uncertainty in that environment, which I will always be grateful.

My grandfather and both of my uncles were supportive of my intellectual curiosity and educational pursuits. We would go out on half-day long walking trips around

downtown Ankara, during which they would patiently address my barrage of questions about what was going on in life.

My dear family, this is your achievement as much as it is mine and words cannot possibly express my gratitude for you.

Table of Contents

List of Figures

List of Tables

Chapter 1. Introduction

1.1 Context and Motivation

Sociotechnical systems (STSs) are complex systems that rely on the successful collaboration between human decision-makers and autonomous technologies to achieve their objectives (Mumford 2006). Historically, STSs have been large in terms of size, governed by hierarchical organizations, and often utilized to provide safety-critical services (O'Sullivan and Sheffrin 2007), e.g., the management of nuclear power plants or air traffic. In the recent years, advances in autonomous decision-making technologies, increasing popularity of personal decision assistance tools, and the benefits of minimizing human involvement in work processes have spread the adoption of, and increased reliance on, STSs (Heydari et al. 2019). Recent surveys of business leaders worldwide find that 60% of all jobs have at least 30% technically automatable activities, and autonomous decision systems will continue to be an important innovation for businesses in the foreseeable future (McKinsey 2019). Consequently, the research community has started to explore systems-thinking approaches to holistically investigate decision making and human-autonomous technology interdependencies (Leveson 2011; Kleiner et al. 2015).

Wide spread adoption of STSs brings together a significant socioeconomic change along with technical challenges. For the systems engineering community, it is no longer sufficient to model and design for the preferences of end-users when making system design decisions (Rich 1983; Hazelrigg 1998). Instead, the end-users, who are independent decision-makers by definition, need to be considered as elements of the STS (De Bruijn and Herder 2009). This is in part due to the growing human-automation collaboration that creates emergent properties of STSs. This collaborative performance is dependent on the beliefs of decision-makers that are influenced by contextual[1] and/or environmental factors

[1] Throughout this dissertation, I use the term c*ontextual variables (or factors)* similar to their use in the efficiency measurement literature (Johnson and Kuosmanen 2012). Contextual factors define operational

that vary during operations (de Visser and Parasuraman 2011). Thus, to enable the effective design and management of future STSs, it is necessary to better understand the sociotechnical interdependencies among people, organizations, their preferences towards collaborating with an autonomous system, and how these preferences vary with respect to contextual factors (Chen and Barnes 2014).

I believe that, STS research could greatly benefit from adopting pragmatic strategies that benefit from the strengths of both theoretical and data-driven empirical approaches. More specifically, I consider modern infrastructure management systems as valuable STS cases to learn from because they have been early adopters of advanced automated decision-making technologies (Pachl 2002). Given the highly intertwined complexity of a ST phenomenon, I quote:

"…that all our science, measured against reality, is primitive and childlike -- and yet it is the most precious thing we have."— Albert Einstein.

Thus, this dissertation establishes the connection among the highly fragmented interdisciplinary literature on STSs and naturalistically[2] investigates an operational complex STS. This is the infrastructure management system of INFRABEL, the Belgian National Railroad Company. INFRABEL's network encompasses about 3,600 kilometers of railway lines, 4,000 track-switching points, and 10,000 rail signals that serve more than 4,000 trains per day. Basing my research on an operational STS allowed me to avoid restrictive assumptions that would have been necessary for alternative simulation-based approaches, it provided verification from domain experts, and helped me to identify gaps in the theory of STSs for future research.

---

and/or environmental conditions and practices that influence the investigated process, yet are: (i) uncontrollable by the decision-maker who oversees the operation, (ii) not resources that are consumed by the process, and (iii) not outputs that are generated by the process.

[2] I use the term *naturalistic* similar to its use in the human factors community (Farrington-Darby et al. 2006), to emphasize that the research presented in this dissertation has been conducted in a non-intrusive way; that did not interfere with the daily operational behavior of the investigated decision-makers.

This dissertation brings together management science with systems thinking and presents a mixed-methodss approach that investigates the sociotechnical interdependencies between people and the autonomous systems they collaborate with. I have organized the dissertation in three mutually exclusive essays, each investigating a distinct facet of STS: safe management, collaboration, and efficiency measurement.

Essay 1 investigates the safe management of STSs by studying the amount of work allocated to safety-critical decision makers who collaborate with autonomous systems towards their common objectives. The primary contribution of Essay 1 is the quantification of Rasmussen's workload boundary (Rasmussen 1997; Cook and Rasmussen 2005) in the form of a Pareto-Koopmans frontier (Koopmans 1951; Farrell 1957). Essay 1 incorporates the influence of contextual factors that could originate from the complexity of the work, macro-ergonomic concerns such as fatigue, or the state of the network. It quantifies their aggregate impact on the workload. Results of Essay 1 indicate that, the aggregate impact of contextual factors can contribute up to 60% of the workload. This establishes the linkage between contextual factors and their aggregate impact.

Essay 2, utilizes the approach provided in Essay 1 to investigate, how human decision-makers and their autonomous partners handle work and how contextual influences shape their collaboration. To elaborate, each person has their own collaboration preferences with autonomous systems, given the plethora of ST contextual influences, and these vary drastically. Consequently, data that describe STS behavior contain a high fraction of influential observations that violate many idealized statistical assumptions (e.g., linearity, normality, etc.) (Hampel et al. 2011; Maronna et al. 2019). Essay 2 establishes the linkage between efficiency measurement techniques and machine learning prediction algorithms to explore the individual contribution of sociotechnical factors on the collaboration of people with autonomous systems. Results of Essay 2 indicate that, during low density and complexity operational situations, autonomous systems can successfully cover for their human supervisors with minor manual interventions. Increased difficulty in operational conditions require both human and autonomous agents to increase their workloads mutually, requiring collaborative performance.

Essay 3 focuses on the management of STSs and explores the complementary and contrasting roles of efficiency measurement approaches that are concerned with the influence of contextual factors. Essay 3 provides a comprehensive literature review that summarizes the existing methodologies in the literature, identifies and implements two of the leading approaches on an operational STS, and discusses the managerial value of these methods with practitioners. Essay 3 documents how certain popular approaches in the literature struggle with the complexity of STSs and provides a structured taxonomy of their fundamental assumptions, limitations, mathematical structure, sensitivity to sample size, and their practical usefulness.

Below I discuss the research questions posed and addressed by each essay, along with their broader contributions to the literature.

1.2 Specific Contributions of Each Essay

1.2.1 Essay 1 – Safe Management

Rasmussen argued that (Rasmussen 1997; Cook and Rasmussen 2005) that a safe operation envelope of STSs is delineated by three distinct boundaries of failure: performance, economic, and workload boundaries. Each is managed by an independent decision-maker within the hierarchical organization that governs the system. Essay 1 focuses on the workload boundary, and quantifies it for the lowest level safety critical decision-makers at INFRABEL. Essay 1 assumes that Rasmussen's workload boundary could be estimated by using a Pareto-Koopmans type frontier that can be empirically estimated through a Data Envelopment Analysis (DEA) approach (Farrell 1957; Charnes, Cooper, and Rhodes 1978; Emrouznejad and Yang 2018). DEA is an axiomatic management science technique that integrates microeconomics and operations research to evaluate the relative efficiency of comparable peers. This dissertation leverages the existing DEA literature. Therefore, I provide a brief review of its fundamentals in Appendix A.

One of DEA's fundamental assumptions is that, the investigated processes need to be homogenous in terms of the employed production technologies and the contextual operational factors. Therefore, it is necessary to identify and properly handle the contextual

operational disparities as experienced by the investigated decision-makers. Essay 1 formulates its first research question for this purpose:

RQ1: "Which sociotechnical factors influence the performance environment for human decision makers in safety-critical environments?"

I address this question by conducting an interdisciplinary literature review of contextual factors that are effective in similar, safety-critical operational environments. Once I have identified these influences, Essay 1 poses a second research question:

RQ2: "How to design, validate, and implement a human performance measurement framework that quantifies Rasmussen's workload boundary?"

I address this question by investigating the process on-site, framing the system boundary around the infrastructure control task of safety critical decision-makers, and reducing the literature driven ideal case model to a real-case model that is limited by the available data. Once the set of variables are identified, I address the research question by following a robust multivariate clustering approach (Triantis, Sarayia, and Seaver 2010; Oscar Herrera-Restrepo et al. 2016). This approach not only accounts for the contextual differences among peers but also quantifies the aggregate impact of contextual variables on the computed efficiency scores. Ergo, Essay 1 addresses the following gaps in the literature:

The primary contribution of this paper, and probably this dissertation, is quantifying Rasmussen's workload boundary (Rasmussen 1997; Cook and Rasmussen 2005). This quantification operationalizes a theoretical concept that remained a qualitative idea for over two decades. Additionally, it provides the macro ergonomics community with a holistic, systems thinking perspective that establishes the relationships among ST factors and their aggregate impact on the efficiency performance of STSs (Carayon et al. 2015; Kleiner et al. 2015). Finally, for the efficiency measurement community, Essay 1 documents the first large scale DEA implementation for an operational infrastructure management system along with the verification of results with domain experts. This extends the traditional application area of DEA and addresses previous suggestions in the literature (Paradi and Sherman 2014; O. Herrera-Restrepo and Triantis 2018).

1.2.2 Essay 2 – Collaboration

Essay 2 extends the workload measurement idea presented in the first essay, to investigate how work is shared among collaborating humans and their autonomous partners. This is a particularly challenging research task even with the rich datasets that enable this study, due to the specific operational set-up at INFRABEL.

To elaborate, INFRABEL decision-makers have an autonomous decision-making system at their disposal. As reported by their supervisors, they have been instructed to use it "to the extent that they feel comfortable with". While the increasing role of automation helps to minimize human errors, in many instances, dynamic contextual characteristics of the infrastructure network, e.g., traffic complexity, render the use of automation ineffective and require manual interventions to sustain reliable operations. The variation in operational practices introduced by the combination of contextual influences and subjective decision-maker preferences, cause the STS data to contain a high fraction of influential observations. Consequently, the data is hard to interpret, especially for the purposes of revealing context dependent relationships. For this purpose, Essay 2 poses its first research question:

RQ1: "How does the workload distribution between collaborating human and autonomous decision-making systems vary given dynamic operational demands?

Essay 2 addresses this question by opening up the blackbox proposed in the first essay. I differentiate between the tasks handled by the human decision-maker, and those delegated to their autonomous partner, through two mutually exclusive Data Envelopment Analysis (DEA) models. This allows one to quantify the work handled by each human and autonomous agent, yet does not capture how the contextual conditions influence decision maker preferences to arrive at the observed work distribution. Essay 2 poses its second research question to establish this relationship:

RQ2: "What are the revealed Controller preferences regarding the workload delegated to automation, given observed contextual infrastructure network characteristics?"

Essay 2 adopts a unique perspective, and treats the DEA scores computed for the first question, as dependent variables that are predicted through machine-learning algorithms, by using the contextual factors as independent variables. For this purpose, Essay 2

implements various prediction algorithms and examines their feature importance that leads to the relative contribution of each contextual variable. In other words, Essay 2 addresses the second research question, by establishing a novel interface between machine learning and DEA that allows one to interpret the influence of contextual variables. This allows one to obtain much higher accuracy than regression based approaches in the literature (Ray 1988; Lovell et.al. 1994; Simar and Wilson 2007; Banker and Natarajan 2008).

1.2.3 Essay 3 – Comparison of Efficiency Estimation Approaches

Essay 3 focuses on the management of STSs and investigates the applicability of efficiency measurement techniques in the literature. The motivation for Essay 3 is explained with the following. Human decisions govern STSs that are susceptible to contextual influences. Recalling that DEA is an axiomatic method, whose fundamental assumption is preserving comparability or homogeneity among investigated peers, violation of this assumption can lead to indefensible results. Thus, the dependency of STSs on contextual factors necessitates an exploration of the existing efficiency measurement methods.

There are numerous analytical methods that are concerned with efficiency measurement with contextual influences and there is an ongoing debate in the literature regarding their utility (Simar and Wilson 2011; Dai and Kuosmanen 2014; Daraio, Simar, and Wilson 2018). Moreover, practical usefulness of these approaches are subject to sample size and data availability issues (Dyson et al. 2001). Essay 3 provides a review of the literature and identifies two leading methods that are fundamentally different from each other. The multivariate clustering approach (Triantis, Seaver, and Sarayia 2010) and the two-stage methods (Cazals, Florens, and Simar 2002; Daraio and Simar 2005; Simar and Wilson 2007).

Essay 3, implements these two contrasting approaches on a STS management problem, and explores the complementary and contrasting roles of each method, along with feedback from domain experts. More specifically, Essay 3 comprehensively investigates the studied methods in terms of their representations of the transformation process, assumptions, limitations, mathematical structure, and their practical usefulness. Additionally, Essay 3

explores the sensitivity of results to sample size by conducting the study on two different datasets (a universe and its subset).

While simple in nature, results of Essay 3 documents how one can interpret the efficiency performance of a STS from two distinct perspectives, while demonstrating how the collective insight of these perspectives could help to obtain a more refined picture of reality. Essay 3 demonstrates that the leading method in the literature to evaluate efficiency and explain contextual influences may not provide a reasonable explanation of the efficiency performance of ST processes due to a violation of its basic assumptions. From this perspective, results of Essay 3 support many others in the research community (Olesen and Petersen 2009; Bădin, Daraio, and Simar 2010; Daraio, Simar, and Wilson 2018; Banker, Natarajan, and Zhang 2019). On the other hand, Essay 3 illustrates how one can utilize the particular strengths of contrasting approaches in a complementary fashion to overcome such limitations.

References

Bădin, Luiza, Cinzia Daraio, and Léopold Simar. 2010. "Optimal Bandwidth Selection for Conditional Efficiency Measures: A Data-Driven Approach." *European Journal of Operational Research* 201 (2): 633–40. https://doi.org/10.1016/j.ejor.2009.03.038.

Banker, Rajiv D., and Ram Natarajan. 2008. "Evaluating Contextual Variables Affecting Productivity Using Data Envelopment Analysis." *Operations Research* 56 (1): 48–58. http://www.jstor.org.ezproxy.lib.vt.edu/stable/25147166.

Banker, Rajiv, Ram Natarajan, and Daqun Zhang. 2019. "Two-Stage Estimation of the Impact of Contextual Variables in Stochastic Frontier Production Function Models Using Data Envelopment Analysis: Second Stage OLS versus Bootstrap Approaches." *European Journal of Operational Research*, Advances in Data Envelopment Analysis, 278 (2): 368–84. https://doi.org/10.1016/j.ejor.2018.10.050.

Carayon, Pascale, Peter Hancock, Nancy Leveson, Ian Noy, Laerte Sznelwar, and Geert van Hootegem. 2015. "Advancing a Sociotechnical Systems Approach to Workplace Safety – Developing the Conceptual Framework." *Ergonomics* 58 (4): 548–64. https://doi.org/10.1080/00140139.2015.1015623.

Cazals, Catherine, Jean-Pierre Florens, and Léopold Simar. 2002. "Nonparametric Frontier Estimation: A Robust Approach." *Journal of Econometrics* 106 (1): 1–25. https://doi.org/10.1016/S0304-4076(01)00080-X.

Charnes, Abraham, William W. Cooper, and Edwardo Rhodes. 1978. "Measuring the Efficiency of Decision Making Units." *European Journal of Operational Research* 2 (6): 429–444. http://www.sciencedirect.com/science/article/pii/0377221778901388.

Chen, Jessie Y. C., and Michael J. Barnes. 2014. "Human–Agent Teaming for Multirobot Control: A Review of Human Factors Issues." *IEEE Transactions on Human-Machine Systems* 44 (1): 13–29. https://doi.org/10.1109/THMS.2013.2293535.

Cook, R., and J. Rasmussen. 2005. "'Going Solid': A Model of System Dynamics and Consequences for Patient Safety." *BMJ Quality & Safety* 14 (2): 130–34. https://doi.org/10.1136/qshc.2003.009530.

Dai, Xiaofeng, and Timo Kuosmanen. 2014. "Best-Practice Benchmarking Using Clustering Methods: Application to Energy Regulation." *Omega* 42 (1): 179–88. https://doi.org/10.1016/j.omega.2013.05.007.

Daraio, Cinzia, and Léopold Simar. 2005. "Introducing Environmental Variables in Nonparametric Frontier Models: A Probabilistic Approach." *Journal of Productivity Analysis* 24 (1): 93–121. https://doi.org/10.1007/s11123-005-3042-8.

Daraio, Cinzia, Léopold Simar, and Paul W. Wilson. 2018. "Central Limit Theorems for Conditional Efficiency Measures and Tests of the 'Separability' Condition in Non-Parametric, Two-Stage Models of Production." *The Econometrics Journal* 21 (2): 170–91. https://doi.org/10.1111/ectj.12103.

De Bruijn, Hans, and Paulien M. Herder. 2009. "System and Actor Perspectives on Sociotechnical Systems." *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 39 (5): 981–992.

Dyson, R. G., R. Allen, A. S. Camanho, V. V. Podinovski, C. S. Sarrico, and E. A. Shale. 2001. "Pitfalls and Protocols in DEA." *European Journal of Operational Research* 132 (2): 245–59. https://doi.org/10.1016/S0377-2217(00)00149-1.

Emrouznejad, Ali, and Guo-liang Yang. 2018. "A Survey and Analysis of the First 40 Years of Scholarly Literature in DEA: 1978–2016." *Socio-Economic Planning Sciences*, Recent developments on the use of DEA in the public sector, 61 (March): 4–8. https://doi.org/10.1016/j.seps.2017.01.008.

Farrell, M. J. 1957. “The Measurement of Productive Efficiency.” *Journal of the Royal Statistical Society. Series A (General)* 120 (3): 253–90. https://doi.org/10.2307/2343100.

Farrington-Darby, T., John R. Wilson, B. J. Norris, and Theresa Clarke. 2006. “A Naturalistic Study of Railway Controllers.” *Ergonomics* 49 (12–13): 1370–94. https://doi.org/10.1080/00140130600613000.

Hampel, Frank R., Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. 2011. *Robust Statistics: The Approach Based on Influence Functions*. John Wiley & Sons.

Hazelrigg, George. 1998. “A Framework for Decision-Based Engineering Design.” *Journal of Mechanical Design* 120 (4): 653–58. https://doi.org/10.1115/1.2829328.

Herrera-Restrepo, O., and K. Triantis. 2018. “Efficiency-Driven Enterprise Design: A Synthesis of Studies.” *IEEE Transactions on Engineering Management* 65 (3): 363–78. https://doi.org/10.1109/TEM.2018.2795563.

Herrera-Restrepo, Oscar, Konstantinos Triantis, William L. Seaver, Joseph C. Paradi, and Haiyan Zhu. 2016. “Bank Branch Operational Performance: A Robust Multivariate and Clustering Approach.” *Expert Systems with Applications* 50: 107–119. http://www.sciencedirect.com/science/article/pii/S0957417415008271.

Heydari, Babak, Zoe Szajnfarber, Jitesh Panchal, Michel-Alexandre Cardin, Katja Holtta-Otto, Gül E. Kremer, and Wei Chen. 2019. “Special Issue: Analysis and Design of Sociotechnical Systems.” *Journal of Mechanical Design* 141 (11). https://doi.org/10.1115/1.4029150.

Kleiner, Brian M., Lawrence J. Hettinger, David M. DeJoy, Yuang-Hsiang Huang, and Peter E. D. Love. 2015. “Sociotechnical Attributes of Safe and Unsafe Work Systems.” *Ergonomics* 58 (4): 635–49. https://doi.org/10.1080/00140139.2015.1009175.

Koopmans, Tjalling C. 1951. *An Analysis of Production as an Efficient Combination of Activities*. Cowles Commission for Research in Economics. New York: John Wiley & Sons.

Leveson, Nancy G. 2011. “Applying Systems Thinking to Analyze and Learn from Events.” *Safety Science* 49 (1): 55–64. http://www.sciencedirect.com/science/article/pii/S0925753510000068.

Lovell, CA Knox, Lawrence C. Walters, and Lisa L. Wood. 1994. “Stratified Models of Education Production Using Modified DEA and Regression Analysis.” In *Data Envelopment Analysis: Theory, Methodology, and Applications*, 329–351. Springer.

Maronna, Ricardo A., R. Douglas Martin, Victor J. Yohai, and Matías Salibián-Barrera. 2019. *Robust Statistics: Theory and Methods (with R)*. John Wiley & Sons.

McKinsey. 2019. “Global AI Survey: AI Adoption Proves Its Worth, but Few Scale Impact.” https://www.mckinsey.com/featured-insights/artificial-intelligence/global-ai-survey-ai-proves-its-worth-but-few-scale-impact.

Mumford, Enid. 2006. “The Story of Socio-Technical Design: Reflections on Its Successes, Failures and Potential.” *Information Systems Journal* 16 (4): 317–42. https://doi.org/10.1111/j.1365-2575.2006.00221.x.

Olesen, O. B., and N. C. Petersen. 2009. “Target and Technical Efficiency in DEA: Controlling for Environmental Characteristics.” *Journal of Productivity Analysis* 32 (1): 27–40. https://doi.org/10.1007/s11123-009-0133-y.

O’Sullivan, Arthur, and Steven M Sheffrin. 2007. *Economics: Principles in Action*. Boston, MA: Pearson/Prentice Hall.

Pachl, Joern. 2002. *Railway Operation and Control*. Mountlake Terrace, WA: VTD Rail Publishing.

Paradi, Joseph C., and H. David Sherman. 2014. “Seeking Greater Practitioner and Managerial Use of DEA for Benchmarking.” *Data Envelopment Analysis Journal* 1 (1): 29–55. https://www.researchgate.net/profile/Joseph_Paradi/publication/281010454_Seeking_Greater_Practitioner_and_Managerial_Use_of_DEA_for_Benchmarking/links/567eb75d08ae051f9ae655de.pdf.

Rasmussen, Jens. 1997. “Risk Management in a Dynamic Society: A Modelling Problem.” *Safety Science* 27 (2): 183–213. https://doi.org/10.1016/S0925-7535(97)00052-0.

Ray, Subhash C. 1988. “Data Envelopment Analysis, Nondiscretionary Inputs and Efficiency: An Alternative Interpretation.” *Socio-Economic Planning Sciences* 22 (4): 167–76. https://doi.org/10.1016/0038-0121(88)90003-1.

Rich, Elaine. 1983. “Users Are Individuals: Individualizing User Models.” *International Journal of Man-Machine Studies* 18 (3): 199–214. https://doi.org/10.1016/S0020-7373(83)80007-8.

Simar, Léopold, and Paul W. Wilson. 2007. “Estimation and Inference in Two-Stage, Semi-Parametric Models of Production Processes.” *Journal of Econometrics* 136 (1): 31–64. https://doi.org/10.1016/j.jeconom.2005.07.009.

———. 2011. “Two-Stage DEA: Caveat Emptor.” *Journal of Productivity Analysis* 36 (2): 205. https://doi.org/10.1007/s11123-011-0230-6.

Triantis, Konstantinos, Devang Sarayia, and Bill Seaver. 2010. “Using Multivariate Methods to Incorporate Environmental Variables for Local and Global Efficiency Performance Analysis.” *INFOR: Information Systems and Operational Research* 48 (1): 39–52. https://doi.org/10.3138/infor.48.1.039.

Visser, Ewart de, and Raja Parasuraman. 2011. “Adaptive Aiding of Human-Robot Teaming: Effects of Imperfect Automation on Performance, Trust, and Workload.” *Journal of Cognitive Engineering and Decision Making* 5 (2): 209–31. https://doi.org/10.1177/1555343411410160.

Chapter 2. Estimation of the Workload Boundary in Sociotechnical Infrastructure Management Systems: the Case of Belgian Railroads

Taylan G. Topcu, Konstantinos Triantis, and Bart Roets

Abstract

Infrastructure systems are large-scale complex socio-technical systems that rely on humans for their safety critical decision-making activities. In the case of railroad networks, hierarchical organizations denoted as traffic control centers (TCCs) operate 24/7 in order to maintain successful network operations. Interacting social and technical factors influence TCC operational environments and thus the overall performance of the railroad system. This research presents a novel data envelopment analysis (DEA) application along with its implementation and validation by investigating the workload boundary of human performance through a case study built for the Belgian railway (INFRABEL) TCCs. We pursue two research foci. The first is to identify organizational, socio-economic, and technical factors that describe the performance environments in which TCC personnel operate. We use these factors to determine relatively homogeneous performance environments using multivariate statistical methods. The second focus is to design and implement on-site a socio-technical performance measurement framework, based on a new and unique dataset at the workstation level that is capable of considering socio-technical heterogeneity. Our approach consists of three steps. First, we apply a two-stage clustering approach to generate statistically relatively homogeneous groups. Second, we calculate meta - and in-cluster efficiency scores. Finally, we assess the validity of our results with INFRABEL. Results reveal three insights: (i) efficiency improvement strategies require further investigation based on temporal trends; (ii) disregarding performance environment heterogeneity leads to over estimation in target setting; and (iii) socio-technical system design could be informed by applying DEA, provided that, domain specific expertise is used in the model formulation.

Keywords: Data envelopment analysis (DEA), socio-technical systems, performance measurement, meta-frontier, infrastructure systems engineering.

Please cite this article as: Topcu, Taylan G., Konstantinos Triantis, and Bart Roets. 2019. "Estimation of the Workload Boundary in Socio-Technical Infrastructure Management Systems: The Case of Belgian Railroads." European Journal of Operational Research 278 (1): 314–29. https://doi.org/10.1016/j.ejor.2019.04.009.

2.1 Introduction

2.1.1 Context and Objective

The focus of this research is the measurement and improvement of efficiency performance of socio-technical systems (STS). We consider efficiency measurement and DEA in particular, as an unifying approach to quantify the states of a safety-critical complex socio-technical system (Rasmussen, 1997; Cook & Rasmussen, 2005). In Rasmussen's system safety model, a widely accepted concept for assessing safety-critical performance, a 'safety envelope' bounds the operating states of a complex system. Three performance frontiers or boundaries define the 'safety envelope'. The 'economic boundary' reflects the minimum economic performance of the system to remain viable. Whereas, the 'safety boundary' represents the risk levels beyond which the system will functionally fail. Finally, the 'workload boundary' accounts for the total amount of work, the system (e.g., a control room) can handle. The objective of this paper is to demonstrate the capability of DEA to compute the workload boundary in a safety critical system for human decision makers, while explicitly taking into account its socio-technical characteristics. The highly disaggregated nature of the data that we have at our disposal allows us to examine this workload boundary specifically for the control-room functions. Although the DEA-computed boundary is of a relative nature, it provides valuable managerial insights for temporal and spatial workload variations (e.g., hourly changes of a workstation). As such, our framework translates Rasmussen's safety model, which is mainly of a descriptive nature (Cook and Rasmussen, 2005), into a quantitative model with significant real-world relevance and extensive opportunities for practitioner feedback towards model improvement.

Based on the need to measure human performance in STSs and within their contextual performance environment, this paper pursues two research questions while presenting a real and unique application of DEA. The first is as follows: "What are the socio-technical

factors that influence the performance environment for human decision makers in safety-critical environments?" We investigate this question through an approach that identifies organizational diagnostics based on a comprehensive literature review of STSs. The second question we investigate is "How to design, validate, and implement a human performance measurement framework that quantifies Rasmussen's workload boundary?" We address this question by providing a detailed walkthrough of a statistically rigorous measurement framework along with the associated validation efforts.

2.1.2 DEA Application

Within this context, and through a unique academia-industry collaboration, we present a novel DEA application paper based on the following criteria:

(i) The application domain: Our research demonstrates how to apply and implement DEA to an infrastructure management system of noteworthy complexity. The implementation of efficiency performance to complex systems remains an open domain of inquiry (Paradi & Sherman, 2014; Triantis, 2015).

(ii) Innovation and the thoroughness of the methodology: This paper provides a multi-disciplinary methodology linking three research domains, i.e., economic production theory, socio-technical systems, and enterprise design (Herrera-Restrepo & Triantis, 2018). We analyze the production transformation processes to identify the socio-technical variables that define the performance environment and differentiate them from the input/output variables (Carayon et al., 2015; Kleiner, Hettinger, DeJoy, Huang, & Love, 2015). Furthermore, we investigate the environmental homogeneity assumption through robust statistical methods (Hubert, Rousseeuw, & Branden, 2005) and ensure the assumption is satisfied through clustering techniques (Hartigan & Wong, 1979; Wong & Lane, 1983). Finally, we quantify the workload boundary through DEA (Rasmussen, 1997; Cook & Rasmussen, 2005).

(iii) The uniqueness and completeness of the used data: This study uses a unique socio-technical dataset obtained from a custom built measurement tool that provides disaggregate measurements of a large-scale infrastructure management system.

(iv) The validation of the proposed methodology and obtained results: Our research is enabled by a unique collaboration of academics and practitioners. It includes guidelines for the design of a performance measurement framework for practical use, its validation by domain experts, and implementation of the developed framework in the field.

(v) Contribution to modeling approaches: For the first time in the literature, we use DEA to incorporate social-technical variables to represent performance environments and to quantify the workload boundary of STSs (Rasmussen, 1997; Cook & Rasmussen, 2005).

(vi) Policy insights for infrastructure management systems: Results of the study document that complexity is the primary driver of risks associated with infrastructure system performance and disregarding social considerations results in unreasonable treatment of human performance.

2.1.3 Terms, Assumptions and Framing

We begin by elaborating on concepts and terms that we use throughout this document. The term "sociotechnical system" (STS) is used to define systems that rely on a "collaboration of humans" to fulfill its mission (Mumford, 2006). In other words, we use the term STS to represent systems with a social subsystem that performs through collaborative efforts of its constituents and governed by organizational rules. Organizational rules in this context are used to define where the control boundaries are drawn and how communication/collaboration is coordinated.

The reliance on human performance in the case of STSs is not an arbitrary design decision but it is inevitable due to how the system functions, which adds to its complexity[3]. For example, in infrastructure management systems, humans conduct arguably the most complex function of the system: the safety critical decision-making activities. These are situations that could potentially lead to loss of life, property, and/or degradation of system performance (Department of Defense, 2012). Safety critical conditions are hard to predict before their occurrence and typically, they are unique. Decision-making under these

[3] We use the term *complexity* to represent system behaviors that result from being composed of many interrelated and interacting elements.

circumstances requires precise situation assessment and immediate reaction. The caveat is that both "assessment" and "reaction" are human decision-making activities that are susceptible to the environment in which the decisions take place. We define the term *performance environment* as the operation space where performance occurs. Two elements shape this space, i.e., internal and external. Internal elements originate primarily from the performer (DMU) and its interaction with the employed production technology. We consider external factors as contextual/environmental Z variables as we know them in the efficiency literature. The external factors significantly affect performance yet are not produced or consumed by the transformation process and are uncontrollable by the performer.

More formally, we define the socio-technical production technology as:

$$ST = \{(\mathbf{x}, \mathbf{h}, \mathbf{y}) \mid \mathbf{x} \text{ and } \mathbf{h} \text{ give rise to } \mathbf{y}, \text{ given } \mathbf{z}\} \quad (1)$$

Here, $x \in \Re_+^p$ represents a vector of inputs of dimension (1 x p), $h \in \Re_+^q$ a vector of decisions of dimension (1 x q), $z \in \Re_+^r$ a vector of socio-technical environmental variables of dimension (1 x r), and $y \in \Re_+^s$ a vector of outputs/outcomes of dimension (1 x s).

For safety-critical systems, leaning on Rasmussen's safety envelope concept (1997, 2005), we delineate this socio-technical operating space by three distinct boundaries or frontiers: the workload boundary ($\partial\Phi$), the economic boundary ($\partial\Psi$), and the safety boundary ($\partial\Sigma$). In this paper, we will estimate the workload boundary $\partial\widehat{\Phi}$, while explicitly accounting for environmental heterogeneity, by executing a two-step clustering DEA approach. Observations on the boundary ($\partial\widehat{\Phi}$) are considered as being the limit of manageable workload, and receive a "workload score" of one. We next revisit one of the core assumptions of DEA and its relationship to the performance environment in this study.

A foundational assumption of DEA (Charnes, Cooper, & Rhodes, 1978) is the assumption that the DMUs operate in relatively homogeneous environments. The homogeneity assumption states that evaluated DMUs have to be "comparable" in terms their input-output specifications, the production technologies they employ, and the environments in which they operate. However, in many DEA studies, DMUs operate in heterogeneous performance environments and do not adhere with this core assumption

(Kuosmanen, Keshvari, & Matin, 2015). In this paper, given the criticality of safety in the application domain, we do not take for granted the comparability assumption and we assure for it rigorously using robust multivariate methods. At this junction, we introduce our infrastructure collaborator.

2.1.4 Our Research Collaborator

INFRABEL is a government owned company whose main responsibilities are building, maintaining, and operating the Belgian railway network. INFRABEL employs close to 12,000 employees and assures the real-time traffic control of some 4,200 trains that serve approximately 800k passengers per day. INFRABEL maintains such a high volume of service by managing a control network of comparable size that consists of over 10k traffic signals and 4k track switch mechanisms. Traffic Control Centers (TCCs) are the centralized locations that perform control and management activities of the network. INFRABEL manages the TCC[4]s by hierarchical organizations that operate 24/7 in eight-hour shifts. Controllers are the TCC personnel tasked with monitoring the state of the infrastructure and interfering through infrastructure control decisions. Controllers perform by managing workstations that provide them with live information regarding the state of the infrastructure and each workstation has a dedicated portion of the physical railroad that it controls. Decisions made by Controllers only physically affect their dedicated control area, yet on an aggregate scale, overall infrastructure performance emerges by the aggregation of the Controller decisions.

INFRABEL previously developed multiple mechanisms to evaluate performance despite the production technology has changed drastically. A DEA model was developed for non-computerized TCCs (Roets & Christiaens, 2015) indicating variations in the overall TCC performance with respect to social and technical variables such as the day of performance and the complexity of the managed tracks. Furthermore, a Controller fatigue estimation study was conducted and it indicates a correlation between fatigue and the tendency to make mistakes (Folkard, Robertson, & Spencer, 2007; Roets & Christiaens,

[4] In the rest of this paper, we use the term TCC to denote the entity composed by the organization that operates the facility along with the infrastructure management equipment on site.

2017). A third study developed a benchmarking model that evaluates the staff efficiency of computerized TCCs, and examined the potential relationship between efficiency levels and human error occurrence (Roets, Verschelde, & Christiaens, 2018). However, Roets, et al. (2018) used purely technical variables limited to entire TCC teams instead of individual Controller performance. Focusing on individual Controller performance offers a significant point of departure for this paper. All previous studies at INFRABEL indicate that performance is an emergent property of the system that needs a holistic investigation that incorporates both social and technical aspects.

TCCs are currently undergoing a major overhaul that significantly changes their production technologies. There has been a recent managerial intervention that created new Controller roles dedicated to safety critical activities (Safety Controller (SC)) and management of traffic (Traffic Controller (TC)). In addition to that, an automated train route-setting device recently implemented provides an "autopilot" feature for non-safety critical decisions. INFRABEL has already implemented these changes in a pilot TCC for test purposes. INFRABEL management expressed the need for detailed performance metrics, capable of assessing Controller workload while accounting for socio-technical performance shaping conditions (e.g., by accounting for fatigue levels). In order to monitor the effect of all changes, INFRABEL has developed a custom-built Business Intelligence (BI) tool to provide measurements of Controller activities at the workstation level. This measurement framework provides a unique socio-technical dataset that is composed of highly disaggregate observations (hourly) of Controller activities under real and highly dynamic circumstances. This unique socio-technical dataset enables this research paper.

2.2 Background

Various bodies of literature contribute to the research of efficiency assessment of socio-technical infrastructure management systems. In this section, we present previous related research in each of these domains along with a discussion on how their collective insights influenced this research.

2.2.1 Sociotechnical Systems (STSs)

The study of STSs has been an area of interest for the communities of human factors and macro-ergonomics, since the focus of these disciplines is human behavior and performance in real contextual environments (Wilson, 2000). The socio-technical approach considers human performance along with its social and technical interactions with the rest of the system. These interactions affect the system's characteristics to a considerable extent (Kroes, 2002; Kroes, Franssen, Poel, & Ottens, 2006). In order to better understand and holistically address the multi-disciplinary nature of STSs, a call to incorporate systems thinking into socio-technical system design has been made (Leveson, 2011) and some researchers from the systems engineering optimization community already approached this domain. However, it is observed that systems engineering driven research still considers the human element as a mechanism to inform technical design decisions (Baxter & Sommerville, 2011; Topcu & Mesmer, 2018). In other words, their approach disregards the interactions between social and technical considerations despite significant evidence indicating that these interactions lead to emergent system properties. For example, risk is known to have bilateral propagation characteristics, meaning that risks originating from social aspects affect technical aspects of the system and vice versa (Wallace, Keil, & Rai, 2004). Performance, is argued to be strictly driven by the interaction of social and technical elements (Kleiner et al., 2015). These assertions emphasize the importance of identifying and capturing socio-technical attributes that create system properties through their dynamic interactions (Carayon et al., 2015).

2.2.2 Safety Critical Sociotechnical Attributes and Their Influence on Employee Performance

There are multiple articles investigating the role of social attributes (e.g., fatigue, management style, supervisor support, cultural traits, stress, etc.) on the overall efficacy of systems. We assume that social attributes shape individual employee performance environments and identify those that we consider closely related to system safety. Two sources: the employee and the organization influence these attributes. We start our discussion with the organization.

Workplace safety is influenced by leadership style (Barling, Loughlin, & Kelloway, 2002) and supervisor support (Hofmann & Morgeson, 1999). Organizational cultural traits have direct implications on system accident rates (Gorman, Cooke, Salas, & Strauch, 2010) especially when two conflicting subcultures are required to collaborate to sustain system activities (Hodgson, Siemieniuch, & Hubbard, 2013). Organizational characteristics have a second order effect on employee performance environments by influencing individual psychological factors. Mental stress is documented to be tightly coupled to system safety (Beehr, 2014). Excessive mental workload increases system failure rates and accidents, especially when it is combined with family-work conflicts (Cullen & Hammer, 2007). In addition to that, the feeling of job insecurity leads employees to disregard safety issues (Probst & Brubaker, 2001). We consider that these attributes are relatively easier to control by the organization and move to individual factors.

An employee sourced attribute related to the Controller performance environment is fatigue. Fatigue is one of the leading causes of human related accidents in the railroad industry (Sussman & Coplen, 2000). In the operational context of railroad Controllers, additional variables reinforce fatigue. The routine of working in shifts disturbs sleep patterns, which in turn result in increased fatigue due to reduced alertness and vigilance (Ferguson, Lamond, Kandelaars, Jay, & Dawson, 2008). Fatigue is observed to be driven mainly by the amount of sleep and the time spent at work (Dawson & McCulloch, 2005). Sleep deprivation was observed to hinder cognitive performance similar to the effect of alcohol intoxication (Van Dongen, Maislin, Mullington, & Dinges, 2003). A study of rail industry employees revealed that mental workload also drives fatigue. However, compared to the lack of sleep it is considered to be of secondary importance (Rosa, 1995; Dorrian, Baulk, & Dawson, 2011).

Another piece of the literature we consider related to human performance is cognitive performance thus we include a brief discussion. Cognitive performance was measured as a function of the familiarity with the given task, the complexity of decision-making rules, and the provision of visual cues (Rubinstein, Meyer, & Evans, 2001). This study concluded that alternating between decision-making tasks with different task familiarity leads to increased costs. The concept of task familiarity brings out an interesting aspect that is

related to the TCC Controller roles, i.e., prospective memory, which is defined as the ability to remember intended actions in the future (Kerns, 2000). For the medicine and aviation sectors, a list of contingency actions to avoid prospective memory related errors was found based on empirical studies (Dismukes, 2012). This study concluded that the distribution of cognitive tasks among human and nonhuman elements is linked to prospective memory capacities. A follow up study argued that a systems driven approach was necessary to fully understand the underlying mechanisms that determine the relationship between memory, performance, and safety (Grundgeiger, Sanderson, & Dismukes, 2015). We concur and would like to add that a systems thinking approach would utilize identified sociotechnical relationships to design better systems. For example, a study conducted in a simulated air traffic control environment has shown that altering the visual decision making cue interface (to support prospective memory) reduces operator errors 11% to 34% (Loft, Smith, & Remington, 2013).

2.2.3 Operational/Contextual Heterogeneity and DEA

DEA is a normative technique that evaluates the relative productive efficiencies of DMUs. The emphasis on "relative" implies that both the employed production technologies and the individual environments where DMUs operate, also known as the operational environments, are comparable. The issue is that there is no universal definition of comparability in the literature. The term "contextual (Z) variables" is widely used in the literature to define variables that are related to performance from both operational and organizational perspectives yet are not directly included in the production technology (Banker & Natarajan, 2008). Related to this point, there is the underlying argument that the contextual variables are "separable" from the production technology. We use the term "performance environment" to define the combination of factors that significantly influence performance yet are not produced or consumed by the DMU. We consider that statistically significant differences in performance environments result in operational/contextual heterogeneity and thus we need to address rigorously.

While there are multiple approaches in the literature that model contextual variables (Banker & Morey, 1986; Ray, 1988; Daraio & Simar, 2005; Simar & Wilson, 2007, 2011; Johnson & Kuosmanen, 2011), we choose an approach that relies on multivariate methods

(Triantis, Sarayia, & Seaver, 2010). We use robust principal component analysis (ROBPCA) that reduces the number of dimensions in the data to hypothetical hyperplanes as an outlier detection method. We use ROBPCA scores to formulate clusters representing relatively homogenous performance subsets. Efficiency analysis then is performed within these clusters (Herrera-Restrepo, Triantis, Seaver, Paradi, & Zhu, 2016). We then compare in-cluster efficiency scores with the meta-frontier efficiency scores to investigate the influence of incorporating performance environments on efficiency designations (O'Donnell, Rao, & Battese, 2008).

2.2.4 DEA Applications in Safety Critical Environments and Infrastructure Management Systems

Despite its academic popularity, applications of DEA to problems of real complexity have been mostly limited to financial and non-profit sectors with very few papers documenting actual implementation and validation[5] (Paradi & Sherman, 2014). We consider two additional criteria to those listed in the first section for the validity of DEA application papers: (i) preserving the axioms of DEA and (ii) compatibility with host organization's preferences. The first point is quite straightforward. DEA is a normative method and failure to comply with its axioms simply provides indefensible results. The second point is arguably the leading factor that limits industry adoption of DEA. To elaborate, organizations not only have unique characteristics in terms of their production technologies and input-output specifications, they also have subjective preferences of the decision-makers and contextual considerations. These considerations introduce additional factors into DEA formulations or could force the analyst to frame the model around measurement-data limitations. An extensive demonstration of how to formulate and implement actual large scale performance evaluation frameworks is provided elsewhere (Borja & Triantis, 2007; Medina-Borja, Pasupathy, & Triantis, 2007; Medina-Borja & Triantis, 2014). Throughout the methodology section of this paper, we will highlight and

[5] We use the term "validation" analogous to its use in the systems engineering terminology (Buede & Miller, 2016). We consider a system validated once the user accepts that the system satisfies its user needs.

discuss how organizational realities of INFRABEL shaped the performance-measurement framework construction.

Applications of DEA in safety critical environments are scarce. A recent study investigated decision making of natural gas transmission plant operators (Azadeh, Gaeini, Motevali Haghighi, & Nasirian, 2016). This study emphasized the importance of considering the principals of macro-ergonomics (Hendrick, 1995) when investigating any technological combination of human, organization, environment, and machine. Another study of Azadeh et al. (2016) investigates intensive care units in hospitals (Azadeh, Tohidi, Zarrin, Pashapour, & Moghaddam, 2016) however, the study disregards operational/contextual heterogeneity issues. Efficiency and effectiveness of railway performance was investigated through a network DEA application however no social variables were included in the model specification (Yu & Lin, 2008). We investigated two DEA applications of European air navigation service providers where both studies highlighted significant issues when representing the notions of safety and complexity (Arnaldo, Comendador, Barragan, & Pérez, 2014; Ćujić, Jovanović, Savić, & Jakšić, 2015). We observed that the hours allocated to safety related activities were treated as a measure of safety outcomes, indicating that the researchers disregarded the complexity and the quality of delivered safety services along with the effect other contextual variables. Another air traffic control study conducted for Federal Aviation Administration (Kopardekar & Magyarits, 2002) multiplied traffic density and traffic complexity to create an aggregate variable denoted as dynamic density (Laudeman, Shelden, Branstrom, & Brasil, 1998). Finally, an interesting DEA study introduced the concept of operator fatigue that is documented to have a strong relationship with safety (Azadeh, Kolaee, & Sheikhalishahi, 2016).

2.3 Methodology

2.3.1 Organizational Diagnostics, Preferences, and Assumptions

As briefly discussed in the introduction, TCCs control dedicated portions of the railroad and are responsible for managing safety and traffic activities in that area 24/7. INFRABEL organized seven workstations in the pilot TCC according to the recently implemented Controller roles. Traffic Controllers (TCs) who manage non-safety critical traffic related

interventions (such as signaling) use five workstations. The safety Controllers (SCs) who handle safety related interventions (such as planned railroad maintenance activities or in the case of incidents) use the remaining two workstations. Both Controller roles work on three nonintersecting eight hour shifts. Controllers form the hierarchical organization together with the managers dedicated to oversee each specific role activities. We denote these managers as Traffic Supervisor and Safety Supervisor. Supervisors do not interfere with routine Controller operations. However, they intervene when there is a challenging decision to make. A higher-ranking manager, i.e., the Traffic Officer manages the entire TCC. Figure 2-1 provides a view inside a random TCC and a simple organizational diagram of the pilot TCC. The BI tool provides data for both Controller activities for each of the seven workstations dedicated to the new roles on an hourly temporal measurement resolution. We discuss the content of the data in Section 2.3.2.

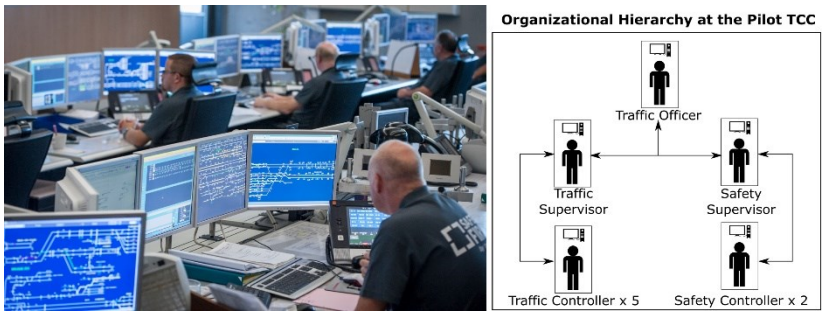

Figure 2-1 Organizational Hierarchy at the Pilot TCC

We argued for the importance of incorporating organizational preferences into the measurement framework formulation, as they are critical for system validation purposes. INFRABEL has two strategic preferences regarding their infrastructure system: (i) the system operates accident free at any cost and (ii) the system minimizes train delays to the lowest possible extent. These preferences have a one to one correspondence with the Controller roles and their decisions, i.e., altering the state of the infrastructure by switching tracks, stopping trains, and so on. There are many Controller decisions, however, these decisions have been aggregated to three groups based on a previous study that evaluated

TCC efficiency at the team level (Roets et al., 2018)[6]. The decision types are movement, adaptation, and safety. Movement decisions correspond to the number (count) of passages at each railroad signal controlled by the Controller and consider the number of train movements as well as local movements known as "shunting". Adaptation decisions represent the changes in traffic flow and include activities such as, merging/splitting trains, re-routing activities, etc. Safety decisions represent procedural interventions to protect track maintenance sites and personnel from incoming trains or other safety measures related to incidents, such as level crossing procedures. Of these decision types, TCs only make movement and adaptation decisions. The TCs recently implemented an automated decision aide system to perform automatically movement decisions. SCs only make adaptation and safety decisions to control the state of the network. However, they constantly monitor the overall train traffic within their dedicated areas to make sure the system is safe.

The **core assumption** that enables this study is that interventions to the state of the system, represented by Controller decisions, implicitly lead to desired INFRABEL outcomes. Therefore, instead of measuring actual INFRABEL outcomes, our approach measures Controller decision outputs and assumes that these outputs lead to the outcomes. In other words, we do not make the distinction between effectiveness and efficiency in this study. Even though Controllers are physically sitting next to each other, their performance environments could be significantly different due to a wide set of factors. These factors include the physical characteristics of the controlled railroad portion, the traffic created by the routes of trains passing through the dedicated area, the human condition of the Controller (e.g., fatigue, stress, experience), and the effect of organizational characteristics on the Controller.

Given this context, the primary need of INFRABEL is a holistic performance measurement framework that (i) incorporates both social (human) and technical

[6] Out of the more than 250 different types of archived Controller decisions, the railway experts selected some 100 decisions as accurately reflecting the decision making process (by avoiding double counting for example). Movement decisions are separated into manual and automatic signal openings, and of the remaining decisions approximately one third was categorized as adapt decisions, the rest as safety decisions.

considerations; (ii) is able to handle highly disaggregate observations; and (iii) is capable of assessing performance under heterogeneity for the new role descriptions. Aside from mentioned performance framework preferences, INFRABEL's domain experts have additional needs related to the implementation and the adoption of the performance measurement tool. These needs are associated with visualizing, monitoring, and interactively analyzing the results. Therefore, we implement the sociotechnical performance framework discussed in this paper at INFRABEL through the BI tool that builds on previous work (Roets et al., 2018). This dimension of our research, i.e., the special attention towards communicating and analyzing the results, is in line with (the few) previous real-world implementations of DEA (Paradi & Schaffnit, 2004; Medina-Borja et al., 2007). This study is also shaped by previous DEA application papers that focus on the visualization and communication of results to practitioners (El-Mahgary & Lahdelma, 1995; Golany & Roll, 1989; Jain, Triantis, & Liu, 2011; Paradi & Sherman, 2014). In the following subsection, we present the data sources of the sociotechnical variables that shape the Controller performance environment.

2.3.2 Model Specification and the Data

Based on the set of assumptions and nature of the transformation process described above, we frame the DMU boundary around one hour of performance of the Controller/Workstation bundle. We use the one hour temporal measurement frame to understand how the workload changes on an hourly resolution. Improved understanding of hourly workload variations could potentially allow INFRABEL to examine flexible Controller schedules instead of fixed eight hour shifts, which could help in fairly balancing the workload over the different controllers.

In order to formulate a framework that considers the multifaceted and sociotechnical nature of performance for the proposed DMU definition, a reductionist approach is applied (Zhao, Triantis, Murray-Tuite, & Edara, 2011). The reductionist approach considers the problem in two stages. First, we specify the model from an ideal point of view where no data availability issues apply. The micro-economic production theory, the literature review, and on-site process/role observations drive the specification. This is fundamentally similar to systems thinking and allows us to identify all variables considered relevant to the

transformation process. The purpose is to ensure completeness and we denote the resulting model as the "ideal case model". Some variables in the ideal case model might not be available by the current measurement system or cannot be measurable with the current state of knowledge and/or expertise. In the second stage of the reductionist approach, the driver is organizational diagnostics, and we reduce the ideal case to fit an attainable abstraction bounded by data availability. We denote the resulting model as the real case model, which is a subset of the ideal case model. The difference between the ideal and the real case model variables indicates valuable sources of information that we do not or cannot measure with current capabilities. Thus, it informs our collaborator about future measurement related investments. Conceptual diagrams of the ideal and real case models in this study considerably leveraged the researcher/railway expert interactions.

In Figure 2-2, we present the ideal case model that was used to initiate face validation efforts during which several INFRABEL experts provided feedback. Horizontal arrows in Figure 2-2 indicate input/output (X and Y) variables, where vertical arrows indicate contextual/environmental (Z) variables that shape the performance environment. The ideal case model reveals that there are four different transformation sub-processes within the TCC describing what the TCs, the SCs, the Traffic Supervisor, and the Safety Supervisor do. In other words, adopting a DEA based lens, each individual role within the TCC form a separate frontier. The ideal case model reveals that the Controllers do not necessarily consume intermediate outputs of the internal Supervisor processes, yet these outputs influence their performance environment. We evaluated the ideal case model presented in Figure 2-2 based on INFRABEL feedback regarding their face validity, data availability restrictions, and organizational preferences.

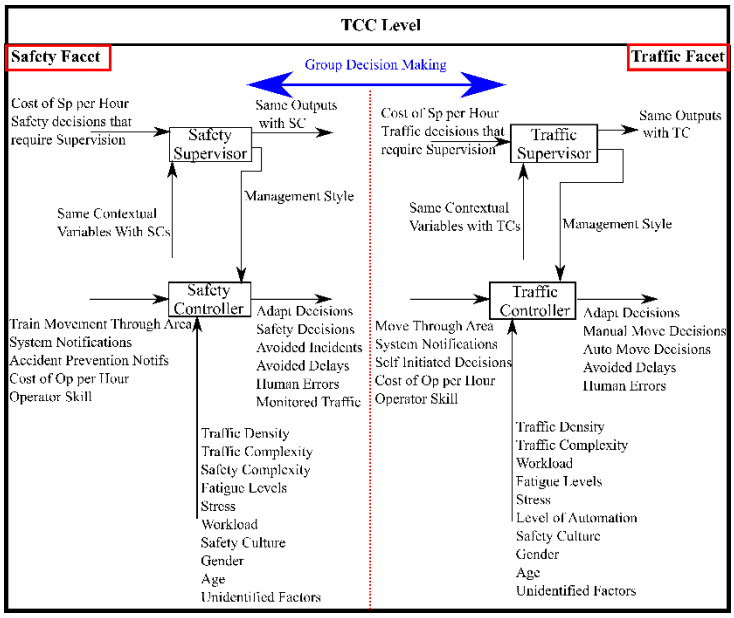

Figure 2-2: The Ideal Case Model

After discussions with the INFRABEL railway experts and based on data availability, we reduce the ideal case model to an attainable real case model[7] that we present in Figure 2-3. We consider the formulation and implementation of the performance measurement framework as an iterative process and use the color-coded variable categorization to establish a shared mental model of our performance measurement model. Employment of research driven visual communication tools continuously assured that railway experts and managers, who have little exposure to the field of performance measurement and mainly DEA, were able to follow/understand the conceptual and modelling basics of the research (Ozbek, de la Garza, & Triantis, 2009). This was crucial for model formulation in terms of receiving relevant and constructive feedback from the organization that acted as a self-verification mechanism. It also established stronger bonds between "outsider" researchers and "insider" practitioners allowing for a more cohesive understanding of railroad management performance.

In Figure 2-3, variables colored with blue have a significant impact on TCC performance. However, it is not possible to obtain a measurement of these variables with the current state of our measurement capabilities. Thus, we considered the blue variables

[7] We provide definitions of the variables included in our analysis in Appendix B.

to be part of the next and future modeling iteration. The blue variables are as follows. Management style, the only different variable for supervisor roles, significantly influences the performance environment of both Controllers as indicated by the literature review. . "Decisions that require Supervisor support" represents the number of times a Controller is unable to make a decision and reaches out for Supervisor support. From an ideal perspective, this variable is the primary input of both Supervisor roles. In addition, it triggers group decision-making activities. Unfortunately, there is no measurement tool in place to monitor its frequency. Therefore, we exclude it from the model. Stress and the safety culture are additional contextual variables that we consider important. Along with the variable denoted as "unidentified factors" (which could be a set of variables that are created by different mechanisms), the measurement of these variables will require separate and focused macro-ergonomics studies. The final blue variable is the group decision making that occurs when a large-scale decision affects multiple control areas. We identify and note importance of this case since it directly contradicts the fundamental DEA assumption of independent DMUs. In TCCs, under specific circumstances, DMUs do interact and collaborate to make a coordinated group decision. We leave this issue for future work.

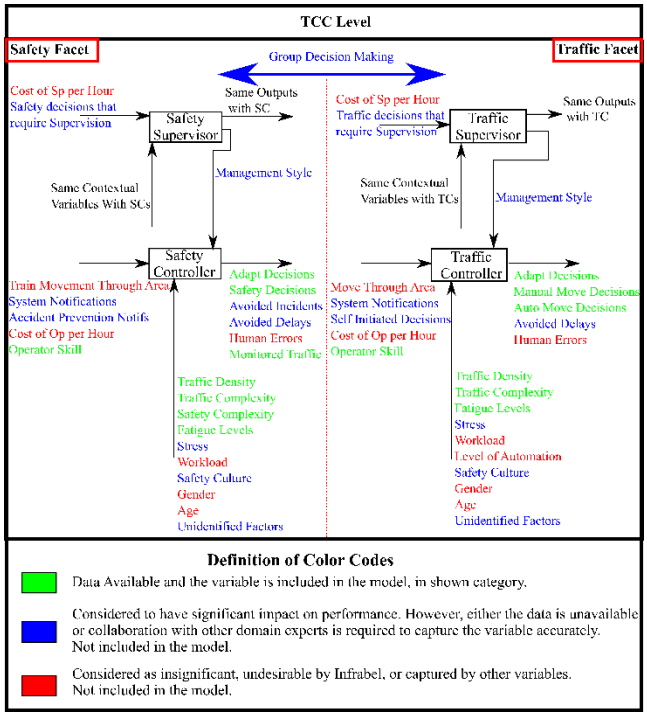

Figure 2-3: The Real Case Model

At this junction, we consider it is appropriate to discuss the data used in this paper. As previously mentioned, a purpose-built BI application, developed at INFRABEL supports our research. BI tool connects work schedule and operational databases, links these data sources at several aggregation levels, prepares the necessary datasets for the performance analysis, and incorporates the intermediate and final results of the research. The dataset includes all green variables as shown in Figure 2-3. As operational IT systems are generally not designed for ex-post performance analysis (Triantis, 2011), one of the biggest challenges faced during the development of the supporting BI tool was to translate the available data into meaningful measures of the traffic control process. In addition, as the workforce and operational systems have different objectives and users (human resource management/scheduling vs. engineering/operations) this required the construction of several additional data tables, linking the original data sources at the desired level of disaggregation. With the current research we significantly extend the dataset created in previous research on computerized railway traffic control centers (Roets et al., 2018), by disaggregating from the team level to the individual Controller level. As such, we created a new and unique sociotechnical database, linking every single traffic control decision not only to the operational circumstances but also and most importantly to the Controller's experience and fatigue levels.

2.3.3 Modeling the Production Process and Rasmussen's Workload Boundary

By applying DEA, we can empirically construct the production possibility frontier or, more specifically the workload boundary that allows us to quantify the workload of DMU. As stated in the introduction, this innovative application of DEA translates Rasmussen's safety envelope, which is of a descriptive nature, into a quantitative (normative) model with significant real-world relevance.

Going back Section 2.3.2., the lack of measurement of blue variables reduces the real case model to mutually exclusive frontier formulations dedicated to both Controller roles. We provide DEA input/output diagram for both Controllers in Figure 2-4. The only input variable for both Controllers is the experience level of the Controller. TC output variables are the number of manual movement decisions, number of automated movement decisions and the number of adaptation decisions. Contextual/environmental Z variables that affect

TC performance environment are traffic density, traffic complexity, and Controller fatigue level. We use an output oriented BCC (Banker, Charnes and Cooper, 1984) model to evaluate efficiency for TC DMUs. Justification of selecting an output oriented variable returns to scale model is the following. Recall that the analyzed data represents the previous safe work statistics (shifts in which zero accidents occurred). We formulate our DEA models based on the assumption that: "given a Controller with x amount of skill level was able to manage y level of actions safely, than a more experienced Controller should be able to handle more workload". We base our assumption on the reality of the transformation process and discussions with domain experts. Our observations of the TCC operation and Controller training data indicate that a more experienced Controller is actually capable of handling larger workloads, given the mitigating effects of the environmental/contextual variables. We understand that this assumption might not hold true for every single DMU and could potentially raise concerns of isotonicity. However, expert feedback indicates that it is ideally correct, thus we consider our fundamental assumption reasonable for the purposes of this exploratory study.

There are two types of outputs for SCs, the number of safety decisions and the number of adaptation decisions. We considered these variables controllable since the SC actively intervenes with the system to produce these safety controls as required by the circumstances. The single uncontrollable output variable is the number of monitored trains highlighted with the red box in Figure 2-4. We consider this variable uncontrollable since the SCs continuously monitor the train traffic in their dedicated area. However, they have no control or approval of any sort over the number of passing trains or how they are handled (within the TCC, this is the TC's responsibility). Therefore, we use an output oriented BCC model with a non-discretionary variable. We use an output-oriented model based on the assumption that an experienced controller should be able to handle larger workloads.

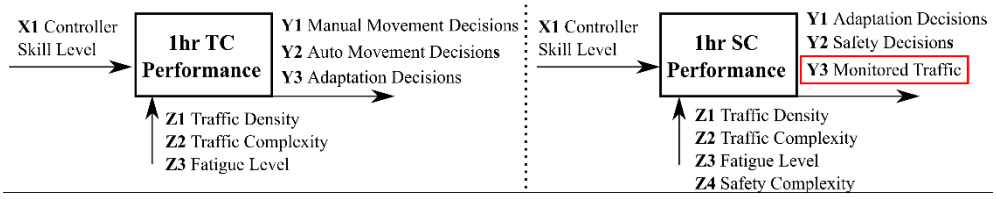

Figure 2-4 Input-Output TC and SC Representations

2.3.4 The Analytical Efficiency Measurement Framework

Both Controller frontiers have multiple contextual/environmental variables that significantly affect and shape their performance environment. Thus, we tailor our framework to handle this heterogeneity. Figure 2-5 provides a diagram summarizing the proposed analytical methodology.

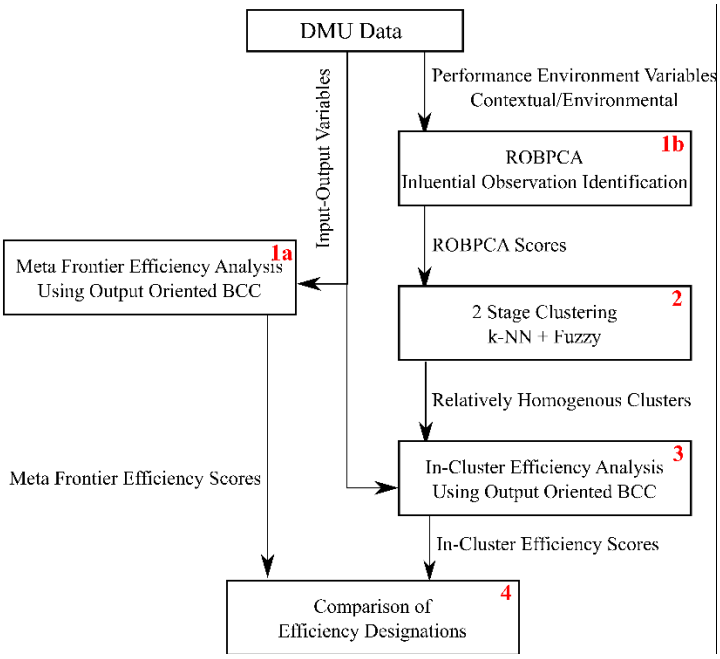

Figure 2-5: Analytical Performance Measurement Framework

Figure 2-5 starts with the unique data obtained from the BI tool. Step 1a is straightforward, input and output variables are extracted for both Controller frontiers and we use an output oriented BCC model (Banker, Charnes, & Cooper, 1984) to calculate meta-frontier efficiency scores. The only difference between controller roles is the non-discretionary output variable, the number of monitored trains, for SCs. In parallel, in Step 1b, only contextual variables are extracted for both frontiers and a ROBPCA is performed (Hubert et al., 2005). Step 1b provides the identification of influential observations, ratio of outliers in the dataset, and principal component scores that define the distance of observations from principal hyperplanes, which we use to reduce the dimensionality in the data. In Step 2, we use the principal component scores to formulate clusters through a two-stage clustering algorithm. First we perform a k-nn nearest neighbor (Wong & Lane, 1983) clustering by considering only two of the nearest observations. Then, we applied a k-means

clustering algorithm (Hartigan & Wong, 1979). Out of all possible number of clusters, we selected the one with minimum jackknife error (Miller, 1974) assuming that it statistically represents the most homogenous performance groups. Step 2 generates relatively homogenous performance subgroups in which performance environments satisfy the comparability assumption. In step 3, we calculate the output oriented BCC efficiency within these clusters by considering only their input and output variables. Finally, we compare the resulting in-cluster and meta-efficiency scores in Step 4.

To summarize and to recall the link with Rasmussen's workload boundary, we propose that Farrell's empirical frontier can be used as a proxy of Rasmussen's workload boundary for human decision makers simply based on the analogy that both represent an attainable limit to performance. Since our data considers only accident free observations, we assume that when a Controller handled a workload successfully in the past, the same Controller can also handle the same workload successfully in the future. We ensure the link of our DEA models with Rasmussen's boundary through variable identification and selection. Thus, we identify the Y variables in parallel with the types of decisions performed by the each Controller. The single X variable represents the only resource used during the transformation process. We select the Z variables from the variables that that the Controller neither consumes nor generates. In other words, they are uncontrollable by the Controller yet they have a strong influence on the transformation process. Therefore, the efficiency scores computed with this model provide a measure of "how efficiently the STS allocates the workload for each controller". As such, our model does not evaluate the individual performance of each Controller, but can be interpreted as evaluating the performance of the STS, in terms of spatiotemporally allocating the workload over the different workstations and Controllers.

The modified output oriented BCC model with non-discretionary outputs used in Step 1a and Step 3 is as follows:

$$\text{Max } \theta \tag{1}$$

$$\text{Subject to} \quad \sum_{j=1}^{n} x_{ij} * \lambda_j \leq x_{ij_0} \;\; \forall i = 1 \; to \; m \tag{2}$$

$$\sum_{j=1}^{n} y_{rj} * \lambda_j \geq \theta * y_{rj_0} \;\; , where \; r \in Discretionary \; Outputs, \forall r = 1 \; to \; s \tag{3}$$

$$\sum_{j=1}^{n} y_{kj} * \lambda_j = y_{kj_0}, where \; k \in Non - Discretionary \; Outputs \tag{4}$$

$$\sum_{j=1}^{n} \lambda_j = 1 \quad (5)$$

$$\lambda_j \geq 0\,, \theta \geq 0 \quad (6)$$

In the set of equations provided above; n represents the number of DMUs, x_{ij} represents the amount of resource i consumed by the j^{th} DMU, λ_j is the weight assigned to a peer j, θ is the efficiency score, $y_{rj}$is the r^{th} output of j^{th} DMU, and y_{rj_0} is the r^{th} output of the DMU under investigation.

Step 1b (Hubert et al., 2005) and Step 2 (Hartigan & Wong, 1979; Wong & Lane, 1983) is described in detail elsewhere. However, we will briefly describe the idea behind the models to ensure completeness and to justify why we consider ROBPCA as a mechanism to ensure DEA's comparability assumption. Step1b focuses on Z variables that define the performance environment, select the best linear combination (principal components) of the variables that contain the most information, and identifies influential observations. We consider that influential observations (outliers) contain the most information regarding the production transformation process and choose to include them in our analysis rather than discard them unless a very large measurement error is found. Identification of influential observations provide a measure of operational heterogeneity since they provide the ratio of outliers to the remaining data in the sample in a robust manner. To this end, we calculate the relative distance of observations from the center of the principal hyperplane. We denote the distances from the principal hyperplane as score and orthogonal distances. Score distance represents how far the observation is when measured parallel from the center of observations on the hyperplane. Orthogonal distance represents how far the observation is when measured vertically from the hyperplane. We record the score and orthogonal distance values for each observation, since they provide a robust and quantified measure of heterogeneity and pass it on to Step 2.

The purpose of Step 2 is to form relatively homogenous performance clusters. To this end we first perform a density based k-nn clustering by using two nearest neighbors (Wong & Lane, 1983) and then group sub-clusters by using a k-means clustering approach (Hartigan & Wong, 1979). To decide on the appropriate number of clusters, we consider the Jackknife error associated with the clustering results and keep iterating until we obtain the lowest error rate. We assume that clusters provide us with relatively homogenous and

comparable subsets of DMUs. Finally, in Step 4 we compare the results of Step1a and Step1b in search for an answer to our research questions, specifically focusing on the relationship between the performance environment and human performance efficiency in STS.

To summarize, as in many other safety-critical and complex systems, increasing pressure for efficiency could also push railway traffic control operations closer to the workload and safety boundaries' (Dekker, 2016). One should pay special attention to the looming dangers of `decrementalism', where small and gradually implemented productivity gains slowly and almost unnoticeably carve out workload and safety margins. The development and implementation of a tool capable of evaluating and monitoring the workload boundary is a first but necessary step in safeguarding traffic control rooms (or any other safety-critical setting) against these progressive performance decrements. Importantly, as highlighted in our literature review, there is an absolute necessity to model sociotechnical systems and their highly heterogeneous performance environments in a multidimensional way. Therefore, in order to assess the Controller workload boundary, we develop a framework that is capable of (i) capturing the multidimensional performance environment and translating it into a discrete number of "contexts" (the clustering phase), and (ii) incorporating the multidimensional nature of the sociotechnical "production process" (the DEA phase). In order to avoid the explicit valuation of each of the dimensions of the sociotechnical system (e.g., put a monetary value on safety, fatigue, or complexity) both phases are based on a relative approach: we assess relative homogeneity in phase 1, and relative efficiency in phase 2.

2.4 Results and Implementation

2.4.1 Test of the Comparability Assumption through Influential Observation Identification – Step 1b

Applying ROBPCA to TC contextual variables yields an orthogonal cutoff distance of 1.6031 and score cutoff distance of 2.7162. Out of 2,919 TC observations, 17 observations are identified as bad leverage points, 128 observations are identified as good leverage points, and 215 observations are identified as orthogonal outliers. We classify the remaining 2,559 observations as regular observations. For SCs, we calculate the orthogonal

cutoff distance measure as 0.8125 and the cutoff score distance as 3.0575. Out of 1031 SC observations, 68 observations are identified as bad leverage points, 99 observations are identified as good leverage points, and 58 observations are identified as orthogonal outliers. We classify the remaining 806 observations as regular observations. We represent the distribution of these influential observations in Figure 2-6.

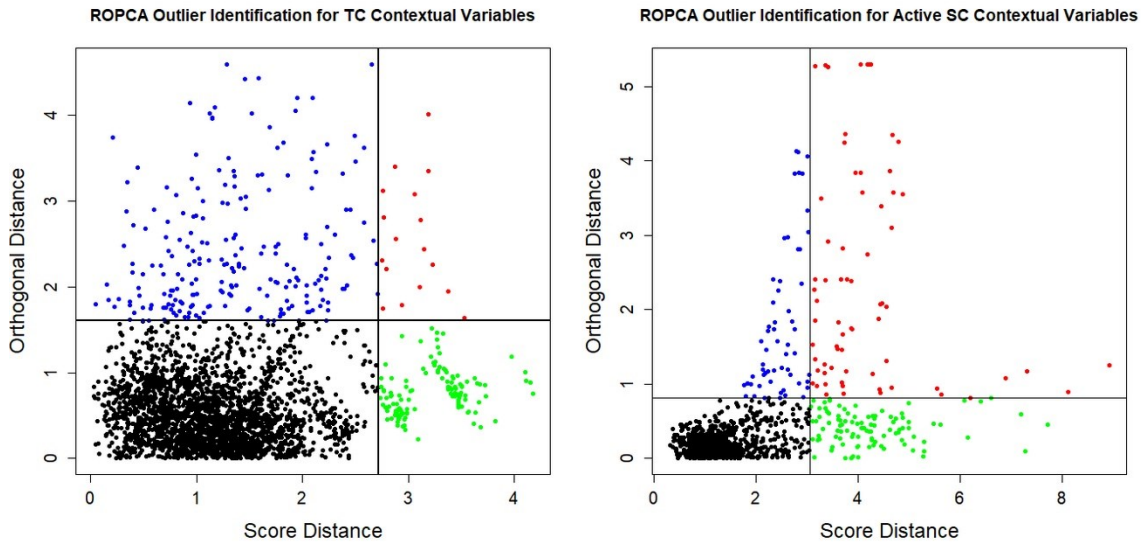

Figure 2-6: Distribution of Influential Observations for TCs and SCs

Results of the ROBPCA reveal that influential observations constitute 21.1% of TC data and 21.8% of SC data. High fractions of influential observations supports our questioning of the comparability assumption. It also suggests that both controller types, experience highly heterogeneous performance environments. The percentage of orthogonal outliers are 18.7% to 5.6 % for TCs and SCs respectively, suggesting high variability in the TC data. We observe the most drastic difference in good and bad influential observations. SCs have 9.6% good leverage points compared to 4.4% of TCs. Bad leverage points demonstrate a larger deficit, as SCs have almost ten times more bad leverage points compared to the TCs. This indicates that the SCs roughly experience ten times more extreme hours and supports the recent INFRABEL decision to create new Controller roles dedicating more skilled personnel to the SC role. In our attempt to formulate relatively homogenous sociotechnical performance subgroups, we store the distances for each controller type and proceed to the next step in our methodology.

2.4.2 2 Stage Clustering – Step 2 Results

We present the jackknife errors for the different number of clusters in Table *2-1* for both TCs and SCs. Even though all investigated clusters demonstrate less that than a 10% Jackknife error, and therefore could be considered appropriate for further analysis (Seaver & Triantis, 1992), we consider that it is more reasonable to use the number of clusters with the minimum error. In the case of TCs, the error rate increased significantly after the three-cluster solution, therefore, we concluded that the three-cluster solution is adequate to represent the data. For SCs, we observed that the two clusters to have the minimum error rate.

Table 2-1 Jackknife Error for the Number of Clusters for both Controllers

	Number of Clusters	**2**	**3**	**4**	**5**
Jackknife Error	**TC**	1.83%	1.60%	2.64%	3.32%
	SC	3.91%	5.95%	7.78%	9.36%

We provide in Table *2-1*, descriptive statistics clusters that reveal interesting insights. The first TC cluster performs under higher traffic density and traffic complexity compared to other TC clusters. Fatigue levels are under the risk level of one and are relatively close for the first and the second TC clusters. The third TC cluster experiences the lowest traffic complexity and density however, it has the highest fatigue level among other TCs. Given the information presented in Table *2-1*, we label TC cluster 1 as high traffic-rested, TC cluster 2 as favorable, and TC cluster 3 as low traffic-fatigued.

Table 2-2 displays the descriptive statistics for SC clusters, revealing how performance environments differ drastically. For SCs, cluster 2 operates with almost twice as much denser traffic. However, the traffic complexity is much higher in SC cluster 1. The first SC cluster has significantly higher fatigue levels with a median above the threshold level of one. Fatigue levels of SC cluster 2 hint that these controllers are well-rested during their shifts. We observe the most significant difference in the safety complexity. As such, SCs in cluster 1 have to make decisions that are almost six times more complex than the second SC cluster. We conclude that the first SC cluster operates in a more demanding operational environment than the second cluster.

Table 2-2 Descriptive Statistics of Controller Clusters

		Score Distance	**Orthogonal Distance**	**Traffic Density**	**Traffic Complexity**	**Fatigue Level**	**Safety Complexity**
TC Cluster 1 n = 1192	Mean	-1.1652	-0.0874	0.8026	3.4167	0.8249	NA
	Median	-1.0942	-0.0399	0.8	2.7	0.7988	NA
	STD Dev	0.5819	0.7672	0.1861	2.4549	0.1193	NA
TC Cluster 2 n = 1267	Mean	0.5546	0.3880	0.2974	1.5852	0.8073	NA
	Median	0.5339	0.4069	0.3067	1.2766	0.7934	NA
	STD Dev	0.5086	0.4625	0.1492	1.3869	0.0822	NA
TC Cluster 3 n = 460	Mean	1.5232	-1.4656	0.2031	0.9223	1.1583	NA
	Median	1.6267	-1.4007	0.1625	0.5395	1.1572	NA
	STD Dev	0.6085	0.6989	0.1687	1.2056	0.1315	NA
SC Cluster 1 n = 231	Mean	3.8237	1.2857	0.4009	1.1310	1.0508	0.2893
	Median	3.6571	0.7235	0.2075	0.7391	1.0775	0.2352
	STD Dev	1.0908	1.3059	0.4870	1.1640	0.2118	0.2580
SC Cluster 2 n = 800	Mean	1.3245	0.2160	1.0954	0.3596	0.8151	0.0563
	Median	1.2408	0.1555	1.1304	0.1886	0.7925	0.0419
	STD Dev	0.5294	0.2130	0.4739	0.4057	0.1048	0.0619

Labeling of the clusters translates the results of the statistical process into an operational language. This proved to be critical for a smooth researcher/railway expert interaction. It is clear that the TC second cluster is operating under the most favorable conditions, however how the sociotechnical performance environments of TC clusters 1 and 3 compare in terms of operationally more demanding conditions remains an open question. We provide a visualization of the observation distribution as measured by their distances to the principal components in Figure 2-7. We retain the clusters as measured by their principal component scores and move on to Step 3.

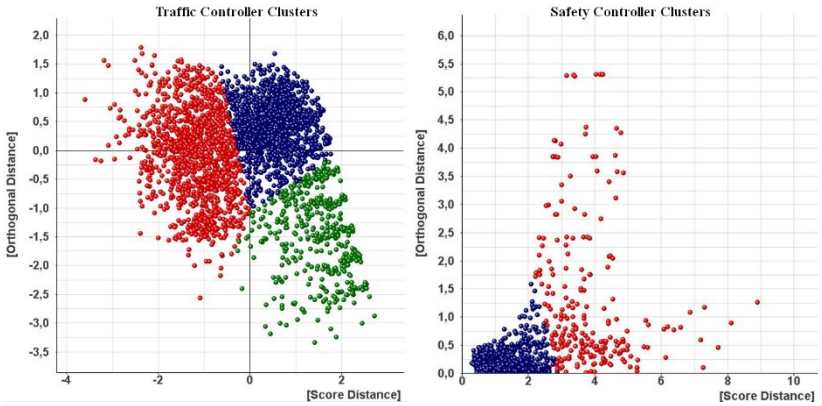

Figure 2-7: Visualization of Clusters on Principal Component Axes

2.4.3 In-Cluster and Meta Frontier Efficiency Analysis – Step 3 & 4

2.4.3.1 TC Performance Analysis

There are 2,919 TC observations in our dataset and we present their output oriented VRS efficiency analysis from both in-cluster and meta-frontier perspectives in Table *2-3*. There are 1,192 observations in the first cluster and this group has the highest average efficiency and minimum variation in in-cluster efficiency distribution. As expected from our TC cluster label designation, this group has higher workload than the second cluster that we considered more favorable prior to the analysis. Following discussions with domain experts at INFRABEL, we consider this expected simply based on the fact that, denser and more complex traffic requires TCs to make more decisions. From this lens, we can interpret the TC efficiency scores as a measure of how busy a TC gets during an hour. The majority of the TC observations belong to the second cluster and its in-cluster efficiency distribution indicate that relatively lower density and complexity simply results in lower workload than the first cluster. Observations in the third cluster are fundamentally different from the first two as they display high fatigue levels even though the traffic density and complexity are significantly low. We observe the effect of low traffic demand on the efficiency score distribution as the third cluster is skewed to the right. This along with higher deviation points to the fact that majority of observations have lower workloads due to low traffic. This also means that we do not capture fatigue necessarily with our definition of workload while concurrently suggesting that INFRABEL does train scheduling by considering Controller fatigue. From a meta-frontier perspective, we observe that all meta efficient DMUs are from cluster 1, as expected. As a general rule of thumb, it is expected for in-

cluster efficiency scores to be higher than meta frontier scores and the difference is denoted as the technology gap (O'Donnell et al., 2008). Interestingly, we observe in Table *2-3* that the mean efficiency score for cluster 3 is lower than the meta-frontier for the entire sample. We attribute this to two factors: (i) drastically low efficiency scores in this cluster as indicated by a strong right skewness and (ii) relatively small size of cluster 3 as its almost one third of the first two clusters. Consequently, the third cluster does not hold enough weight in the calculation of the mean Meta efficiency score.

Table 2-3 TC In-Cluster Efficiency Summary

Estimation of the Workload Boundary	**TC Cluster 1** *Dense and Highly Complex Traffic with Rested Controllers*	**TC Cluster 2** *Medium Density and Complexity with Rested Controller*	**TC Cluster 3** *Fatigued Controller facing Low Density and Complexity*	**TC Meta Frontier** *All DMUs*
Mean Efficiency	0.6483	0.5493	0.3544	0.4082
Median Efficiency	0.6587	0.5744	0.3093	0.3571
Standard Deviation	0.1634	0.2369	0.2675	0.2459
Mean Meta Frontier Efficiency of the Cluster	0.4280	0.4174	0.3381	NA

In order to validate our results, we collaborate with domain experts and focus on the extreme cases in each individual cluster along with a relative comparison of efficiency distribution with respect to other clusters. When we focus on the observations with the lowest efficiency scores in the second and the third clusters (clusters with high standard deviation), we observe that these DMUs exhibit specific traits. Observations in the 2nd cluster are usually during the day, but in-between rush hour traffic peaks (which are mostly captured by cluster 1). The traffic in this cluster generally runs smoother than the peak hours, which leads to less adaptation decisions. In some cases very low adaptation decisions are made, creating the efficiency scores that are at the lower end of this cluster. Observations in the 3rd cluster are usually on the night shift or in low volume traffic areas. This cluster has the lowest number of observations because during the night shift, several workstations are merged together thus less workstations are active. In some extreme cases, the number of incoming trains can still be very low despite the merger of workstations leading to drastically low efficiency scores. Thus, we should not interpret low efficiency

scores as an indicator of low controller performance. Instead, **we should treat the efficiency scores as a quantification of the Controller workload.** The calculation of the workload boundary and the associated workload efficiencies provides management with a tool capable of not only quantifying the manageable workload (the boundary), but also pinpoint temporal and spatial workload issues. A detailed temporal (e.g., hourly) and spatial (e.g., workstation) analysis can shed additional light on the revealed inefficiency patterns, and allow for a more optimized and fair distribution of the workload across the controller population. Moreover, the statistical identification of the sociotechnical clusters provides important additional insights into the temporal and spatial patterns of the performance environment. Possible improvement strategies for optimizing workload patterns can be the merger of workstations with less demanding hours, or the separation of workstations during longer periods of sustained effort and expected fatigue levels.

2.4.3.2 SC Performance Analysis

For SCs, we present the summary of efficiency scores in Table 2-4 by using an output oriented VRS model with an uncontrollable monitored traffic variable. The first cluster demonstrates much higher workload as expected by the performance environment designations. Roughly, 35% of the DMUs in the first cluster have an efficiency score above 0.5. On the other hand, DMUs in the second cluster have a median efficiency score of 0.2151 and only 20% of the DMUs have an efficiency score above 0.5. We can draw a couple of interesting insights from this. The first is our interpretation of efficiency scores in this study. We observe an increase in the productive efficiency scores when environmental conditions get more demanding (e.g., fatigue levels are higher, complexity levels are higher). Considering that each SC intervention is safety-critical (and recalling our model specification), we make the following observation: **when the system and the decisions associated with it becomes more complex, the workload of SCs increase to keep the system accident free, even under increasing controller fatigue.**

We consider it is necessary to compare meta and in-cluster efficiency scores in order to be able to draw more insights and to be able to see how the DMUs considered efficient within the homogenous subset compares against others. Interestingly, we observe that meta-efficient DMUs are coming from both SC clusters. We observe different conditions

to drive the workload for these observations from a meta-frontier perspective. High number of trains that need to be observed drive the meta-efficient observations from the second cluster. In some cases observed around rush hours, the amount of traffic is so high, DMUs are assigned high efficiency scores despite the fact that no safety actions were required. This demonstrates the importance of integrating efficiency analysis with the rest of organizational data, as it significantly makes it easier to access additional operational information.

Table 2-4 SC Efficiency Summary – Estimation of the Workload Boundary

Estimation of the Workload Boundary	**SC Cluster 1** *Low Density, High Complexity Traffic with Fatigued Controller and High Safety Complexity* *More Demanding Performance Environment*	**SC Cluster 2** *High Density and Low Complexity Traffic with Rested Controllers and Low Safety Complexity* *Favorable Performance Environment*	**SC Meta Frontier** *All DMUs*
Average Efficiency	0.4093	0.3094	0.2336
Median Efficiency	0.3834	0.2151	0.1541
Standard Deviation	0.2832	0.2706	0.2276

To visualize and to demonstrate the effect of considering the sociotechnical performance environment, we present hourly efficiency changes on workstations from both the meta and in-cluster perspectives in Figure *2-8*. As a general remark, we observe that in-cluster efficiency scores are always at least as high as the meta-frontier efficiency scores, by definition. The difference in the efficiency levels and the sometimes notably different efficiency patterns highlight the effect of satisfying the core comparability assumption of DEA, as in-cluster scores are much lower than the meta-efficiency scores. We interpret the potential consequences of using meta-efficiency scores instead of in-cluster scores from several standpoints. **From a system safety perspective**, recalling Rasmussen's (1994) safe operation envelope, making staff alignment decisions based on meta-scores could simply push the system outside the safe operation envelope. Considering that railroad accidents are catastrophic, we should avoid this at all costs. **From a managerial perspective,** relying on meta-scores would lead to setting infeasible Controller performance targets. This would not only underestimate the staff performance but more importantly, it would negatively

affect all the safety related attributes we discussed in Section 2.2.2., eventually creating a less-safe infrastructure system. **From a DEA perspective,** it is a clear demonstration of how adhering to core assumptions of the method is crucial. Especially for highly disaggregate cases similar to ours could lead to indefensible and misleading results.

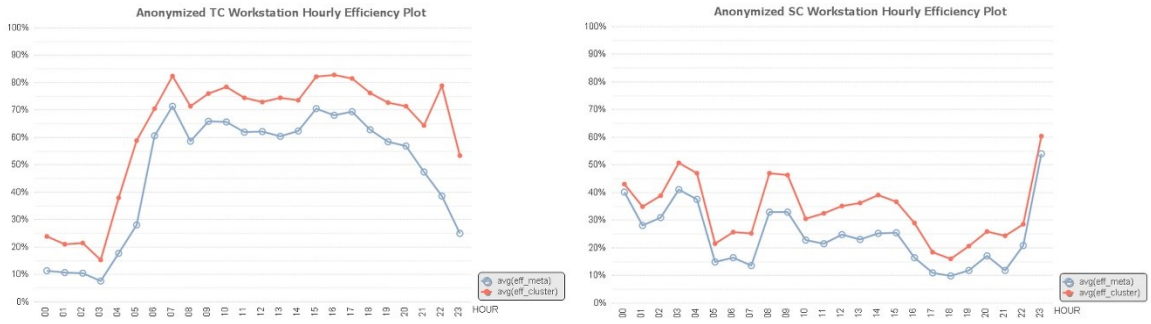

Figure 2-8 Visualization of Meta vs. In-Cluster Efficiency Scores

2.4.4 Validation, Implementation, and Usefulness of Considering Sociotechnical Factors

We carried out the validation efforts by ensuring face validation from INFRABEL regarding the appropriateness of the developed framework in terms of addressing their needs discussed in Section 2.3.1. INFRABEL experts and management evaluated the contextual clustering and relative efficiencies and provided valuable insights on issues or opportunities related to optimal staff alignment. In addition, INFRABEL experts and management paid close attention to the workload boundary as safety is of strategic importance to railway companies and infrastructure managers.

Both consistently low or consistently high[8] efficiencies can lead to a range of issues related to staff well-being, employee morale, absenteeism, and in some cases human error. This is consistent to previous work by Paradi et al. (2014) where best practices for bank branch efficiency are not only positioned on but also close to the estimated production frontier. If bank branch staff is 'pushed to the limit' they can respond by eventually breaking down.

Other examples of possible managerial actions initiated or supported by the efficiency results are as follows. The optimization of team composition (adding or retrieving team

[8] Indicating overstaffing or situations where staffing levels could be close to their workload limits.

members during a work shift); changing shift transition times (e.g., earlier or later start for some of the team members); or negotiating with the asset management department to obtain a more convenient spread of infrastructure maintenance works (a main driver of the "safety" output). Evidently, additional information from field management or staff is necessary before taking appropriate action. However, the efficiency results allow senior management to focus on the most prominent efficiency issues, keep a finger on the pulse by monitoring their evolution, and evaluating the impact of their decisions.

Leveraged by the BI tool, this research received a very positive feedback from both the experts and the management of INFRABEL. The interactive capabilities of the BI application allow for a more detailed analysis of the clustering and efficiency results. We accomplish this at different levels of data granularity, ranging from top-level aggregation to each individual Controller. A key concept underpinning each analysis is our ability to interact with INFRABEL. At any aggregation level and at any moment, users can instantly switch to the highly detailed operational data such as the details of the staff roster, the traffic controller actions, and safety controller actions. This allows business experts to explore the efficiency results in an intuitive yet quantitative way, challenge and complement the findings with their expert knowledge, detailed staff roster and operations data, and feedback from the field. In summary, business experts can extend their analysis beyond the traditional efficiency results, and dive deeper into the possible causes or main drivers of inefficiency.

The objective of the permanent performance measurement is not only to track the efficiency of the aligned traffic control staff, but also - and certainly no less importantly – to identify areas or patterns of possible overload, especially when related to the safety component of the analysis. For example, systematically recurring high efficiencies could point at a structural overload of the staff involved. We could add more months as time progresses. The managerial objective of permanent performance measurement is not only to assess and track the efficiency of the aligned traffic control staff, but also - and certainly no less importantly - identify areas or patterns of potential over- or understaffing, especially when related to safety. For example, systematically recurring high safety weighted efficiencies could point at a structural understaffing of the Controllers involved.

2.5 Conclusions & Future Work

We believe that the actual implementation of our framework is a tangible response to the Paradi and Sherman call for a more active 'selling' of the DEA concept to organizations (Paradi & Sherman, 2014). In contrast to its academic success, DEA has shown an only limited use in practice. Paradi and Sherman suggest augmenting DEA's use, until the 'tipping point' is reached where practitioners and managers recognize for its power and versatility. The suggestions for further development of DEA methodologies and applications (i.e., offering easily accessible DEA software and an increased cooperation between researchers and practitioners) could be key strategic elements in reaching this breakthrough (Liu, Lu, Lu, & Lin, 2013). In addition, we believe that the use of a concomitant software application such as our BI tool can provide a substantial lever for management acceptance. 'Selling' DEA and its results in combination with (or integrated with) an advanced reporting and analysis tool, especially when tailored to the needs and concerns of the management, provides much more added value. Enriching the DEA results with an automated reporting system was also the approach in one of the few other successful DEA deployments, the large-scale implementation at the American Red Cross (Borja & Triantis, 2007; Medina-Borja et al., 2007; Medina-Borja & Triantis, 2014). In this paper, we have taken this approach beyond the reporting aspects alone and added the interactive analysis as a new key component. Information technology has much evolved since Golany and Roll (Golany & Roll, 1989) first suggested 'report generation' and 'graphical data analysis' in their influential DEA application procedure, and the advent of BI software now provides a plethora of functions for exploring and probing efficiency results. The ease of use and interactive ability of these tools empowers management and experts to discover, analyze and monitor efficiency patterns at the click of a mouse, even within very large data sets. In addition, our custom-built application does not only provide value for money in its daily operation as a managerial tool, but also unlocks the full potential of the business experts during the iterative phases of model building and validation. Therefore, we suggest systematically offering the DEA concept as part of a larger and comprehensive solution to the analysis of performance consisting of back-end efficiency calculations and advanced front-end reporting and analysis capabilities. Our successful DEA deployment has proven the applicability of the advocated approach.

The future work associated with this research could involve many different directions given the depth of the study. The first subject is the incorporation of variables colored with blue in the ideal case study. We believe a macro-ergonomics based assessment of TCC would be a fruitful research venture to explore, as the literature indicates, social aspects as workplace culture, management styles, stress, and work-life trends have a direct relationship with human performance. Building on that, we conducted this study assuming that none of the Controllers (DMUs) interacts with each other while we clearly knew from the very beginning this assumption was false. This is part due to missing data to establish the interface between Controllers (DMUs) and the lack of measurement regarding the group decision making that takes place when TCC personnel is faced with a challenging decision. A potential future research opportunity lies in observing the TCC room in real time especially when group decision making occurs and relating that with the existing data measurement framework. Given that measurement regarding macro-ergonomic factors are available, we could model overall TCC performance using a Network DEA approach that considers Controllers and Supervisors as sub processes of the larger TCC performance. Even if the data were available, our preliminary survey of the NDEA literature did not yield any models where intermediate outputs of sub processes constitutes the contextual environment for other sub processes – which is the case in TCCs.

We based our analysis that we performed on an ex-post approach where we evaluated results long after the events leading to performance have already occurred. Considering the dynamic nature of infrastructure systems, decisions based on ex-post analysis might be suboptimal or even not applicable in the future. Therefore, we believe that formulating a dynamic performance assessment framework could provide more insights. Finally, we did not provide a temporal and spatial aggregation of performance. Results of our study clearly indicate the need for aggregation, not only for the duration of shifts, but with the interfacing TCCs to obtain a more holistic understanding of performance.

One of the cornerstones of our study is to dive deep into the details of the production process, by constructing datasets and models at workstation and Controller level. As this level of data disaggregation draws the research nearer to an assessment of individual performance and behavior, a general caveat is however in order. With a frictionless

implementation of the performance measurement system in mind, future research will need to continue its non-intrusive approach, and ensure the personal privacy of the individual traffic controllers. Particularly in highly unionized environments – such as the railways or air traffic control – this is critical in terms of user acceptance and satisfaction. Therefore, to mitigate potential implementation issues, an extension of the research efforts towards the literature on 'Electronic Performance Monitoring' systems - and their effects on organizational and individual performance – may be warranted (see, e.g., (Alge & Hansen, 2014)).

Our research has gone only so far in providing feedback to theory. Given that we view our research as multi-disciplinary, we borrow concepts from other domains without fundamentally changing other domains or reaching the limits of economic production theory. Our next opportunity is to collaborate with researchers in other domains (e.g., human factors, decision theory) where we could undertake a number of theoretical challenges and questions. For example, to what extent do Controller preferences affect the definition of the production possibility set? How do human factor considerations such as distributed situational awareness affect the definition and the consideration of the contextual variables? In the end, how do the answers to these questions challenge the production axioms that we assume in efficiency analysis? In other words, what are the limits of economic production theory and does this application lend itself to explore these limits?

Acknowledgements

We would like to express our gratitude for Dr. Renaat van de Kerkhove, Dr. Alex Fletcher, and Kristof van der Strieckt from INFRABEL for preparing the data and assisting our research with their invaluable feedback. It is however to be noted that the views expressed in this paper are those of the authors and do not necessarily reflect the opinions of INFRABEL. Authors also would like to thank Dhruv Patel and Dr. Oscar Herrera-Restrepo from Virginia Tech for their suggestions and insights regarding our framework.

References

Alge, B. J., & Hansen, S. D. (2014). Workplace monitoring and surveillance research since 1984: A review and agenda. Routledge, New York.

Arnaldo, R. M., Comendador, V. F. G., Barragan, R., & Pérez, L. (2014). European Air Navigation Service Providers' Efficiency Evaluation through Data Envelopment Analysis (DEA). In 29th Congress of the International Council of the Aeronautical Sciences.

Azadeh, A., Gaeini, Z., Motevali Haghighi, S., & Nasirian, B. (2016). A unique adaptive neuro fuzzy inference system for optimum decision making process in a natural gas transmission unit. Journal of Natural Gas Science and Engineering, 34, 472–485. https://doi.org/10.1016/j.jngse.2016.06.053

Azadeh, A., Kolaee, M. H., & Sheikhalishahi, M. (2016). An integrated approach for configuration optimization in a CBM system by considering fatigue effects. The International Journal of Advanced Manufacturing Technology, 86(5–8), 1881–1893. https://doi.org/10.1007/s00170-015-8204-x

Azadeh, A., Tohidi, H., Zarrin, M., Pashapour, S., & Moghaddam, M. (2016). An integrated algorithm for performance optimization of neurosurgical ICUs. Expert Systems with Applications, 43(Supplement C), 142–153. https://doi.org/10.1016/j.eswa.2015.08.042

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

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

Banker, R. D., & Natarajan, R. (2008). Evaluating Contextual Variables Affecting Productivity Using Data Envelopment Analysis. Operations Research, 56(1), 48–58. Retrieved from http://www.jstor.org.ezproxy.lib.vt.edu/stable/25147166

Barling, J., Loughlin, C., & Kelloway, E. K. (2002). Development and test of a model linking safety-specific transformational leadership and occupational safety. Journal of Applied Psychology, 87(3), 488. Retrieved from http://psycnet.apa.org/journals/apl/87/3/488/

Baxter, G., & Sommerville, I. (2011). Socio-technical systems: From design methods to systems engineering. Interacting with Computers, 23(1), 4–17.

Beehr, T. A. (2014). Psychological Stress in the Workplace (Psychology Revivals). Routledge.

Borja, A. M., & Triantis, K. (2007). A conceptual framework to evaluate performance of non-profit social service organisations. International Journal of Technology Management, 37(1/2), 147. https://doi.org/10.1504/IJTM.2007.011808

Buede, D. M., & Miller, W. D. (2016). The Engineering Design of Systems: Models and Methods. John Wiley & Sons.

Carayon, P., Hancock, P., Leveson, N., Noy, I., Sznelwar, L., & Hootegem, G. van. (2015). Advancing a sociotechnical systems approach to workplace safety – developing the conceptual framework. Ergonomics, 58(4), 548–564. https://doi.org/10.1080/00140139.2015.1015623

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

Cook, R., & Rasmussen, J. (2005). "Going solid": a model of system dynamics and consequences for patient safety. BMJ Quality & Safety, 14(2), 130–134. https://doi.org/10.1136/qshc.2003.009530

Ćujić, M., Jovanović, M., Savić, G., & Jakšić, M. L. (2015). Measuring the Efficiency of Air Navigation Services System by Using DEA Method. International Journal for Traffic and Transport Engineering, 5(1).

Cullen, J. C., & Hammer, L. B. (2007). Developing and testing a theoretical model linking work-family conflict to employee safety. Journal of Occupational Health Psychology, 12(3), 266.

Daraio, C., & Simar, L. (2005). Introducing Environmental Variables in Nonparametric Frontier Models: a Probabilistic Approach. Journal of Productivity Analysis, 24(1), 93–121. https://doi.org/10.1007/s11123-005-3042-8

Dawson, D., & McCulloch, K. (2005). Managing fatigue: it's about sleep. Sleep Medicine Reviews, 9(5), 365–380.

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

Department of Defense. (2012). Department of Defense Standard Practice System Safety (No. MIL-STD-882E).

Dismukes, R. K. (2012). Prospective memory in workplace and everyday situations. Current Directions in Psychological Science, 21(4), 215–220.

Dorrian, J., Baulk, S. D., & Dawson, D. (2011). Work hours, workload, sleep and fatigue in Australian Rail Industry employees. Applied Ergonomics, 42(2), 202–209. https://doi.org/10.1016/j.apergo.2010.06.009

El-Mahgary, S., & Lahdelma, R. (1995). Data envelopment analysis: Visualizing the results. European Journal of Operational Research, 83(3), 700–710. https://doi.org/10.1016/0377-2217(94)00303-T

Ferguson, S. A., Lamond, N., Kandelaars, K., Jay, S. M., & Dawson, D. (2008). The impact of short, irregular sleep opportunities at sea on the alertness of marine pilots working extended hours. Chronobiology International, 25(2–3), 399–411.

Folkard, S., Robertson, K. A., & Spencer, M. B. (2007). A Fatigue/Risk index to assess work schedules. Somnologie-Schlafforschung Und Schlafmedizin, 11(3), 177–185.

Golany, B., & Roll, Y. (1989). An application procedure for DEA. Omega, 17(3), 237–250. https://doi.org/10.1016/0305-0483(89)90029-7

Gorman, J. C., Cooke, N. J., Salas, E., & Strauch, B. (2010). Can Cultural Differences Lead to Accidents? Team Cultural Differences and Sociotechnical System Operations. Human Factors, 52(2), 246–263. https://doi.org/10.1177/0018720810362238

Grundgeiger, T., Sanderson, P. M., & Dismukes, R. K. (2015). Prospective memory in complex sociotechnical systems. Zeitschrift Für Psychologie.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics), 28(1), 100–108. https://doi.org/10.2307/2346830

Hendrick, H. W. (1995). Humanizing Re-Engineering for True Organizational Effectiveness: A Macroergonomic Approach. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 39(12), 761–765. https://doi.org/10.1177/154193129503901202

Herrera-Restrepo, O., & Triantis, K. (2018). Efficiency-Driven Enterprise Design: A Synthesis of Studies. IEEE Transactions on Engineering Management, 65(3), 363–378. https://doi.org/10.1109/TEM.2018.2795563

Herrera-Restrepo, Oscar, Triantis, K., Seaver, W. L., Paradi, J. C., & Zhu, H. (2016). Bank branch operational performance: A robust multivariate and clustering approach. Expert Systems with Applications, 50, 107–119.

Hodgson, A., Siemieniuch, C. E., & Hubbard, E.-M. (2013). Culture and the Safety of Complex Automated Sociotechnical Systems. IEEE Transactions on Human-Machine Systems, 43(6), 608–619. https://doi.org/10.1109/THMS.2013.2285048

Hofmann, D. A., & Morgeson, F. P. (1999). Safety-related behavior as a social exchange: The role of perceived organizational support and leader–member exchange. Journal of Applied Psychology, 84(2), 286.

Hubert, M., Rousseeuw, P. J., & Branden, K. V. (2005). ROBPCA: A New Approach to Robust Principal Component Analysis. Technometrics, 47(1), 64–79. https://doi.org/10.1198/004017004000000563

Jain, S., Triantis, K. P., & Liu, S. (2011). Manufacturing performance measurement and target setting: A data envelopment analysis approach. European Journal of Operational Research, 214(3), 616–626. https://doi.org/10.1016/j.ejor.2011.05.028

Johnson, A. L., & Kuosmanen, T. (2011). One-stage estimation of the effects of operational conditions and practices on productive performance: asymptotically normal and efficient, root-n consistent StoNEZD method. Journal of Productivity Analysis, 36(2), 219–230. https://doi.org/10.1007/s11123-011-0231-5

Kerns, K. A. (2000). The CyberCruiser: An investigation of development of prospective memory in children. Journal of the International Neuropsychological Society, 6(1), 62–70.

Kleiner, B. M., Hettinger, L. J., DeJoy, D. M., Huang, Y.-H., & Love, P. E. D. (2015). Sociotechnical attributes of safe and unsafe work systems. Ergonomics, 58(4), 635–649. https://doi.org/10.1080/00140139.2015.1009175

Kopardekar, P., & Magyarits, S. (2002). Dynamic density: measuring and predicting sector complexity [ATC]. In Digital Avionics Systems Conference, 2002. Proceedings. The 21st (Vol. 1, pp. 2C4–2C4). IEEE.

Kroes, P. (2002). Design methodology and the nature of technical artefacts. Design Studies, 23(3), 287–302. https://doi.org/10.1016/S0142-694X(01)00039-4

Kroes, P., Franssen, M., Poel, I. van de, & Ottens, M. (2006). Treating socio-technical systems as engineering systems: some conceptual problems. Systems Research and Behavioral Science, 23(6), 803–814.

Kuosmanen, T., Keshvari, A., & Matin, R. K. (2015). Discrete and Integer Valued Inputs and Outputs in Data Envelopment Analysis. In Data Envelopment Analysis (pp. 67–103). Springer, Boston, MA. https://doi.org/10.1007/978-1-4899-7553-9_4

Laudeman, I. V., Shelden, S. G., Branstrom, R., & Brasil, C. L. (1998). Dynamic density: An air traffic management metric (No. NASA-TM-1998-112226). Moffett Field, California: NASA Ames Research Center.

Leveson, N. G. (2011). Applying systems thinking to analyze and learn from events. Safety Science, 49(1), 55–64.

Liu, J. S., Lu, L. Y. Y., Lu, W.-M., & Lin, B. J. Y. (2013). A survey of DEA applications. Omega, 41(5), 893–902. https://doi.org/10.1016/j.omega.2012.11.004

Loft, S., Smith, R. E., & Remington, R. W. (2013). Minimizing the disruptive effects of prospective memory in simulated air traffic control. Journal of Experimental Psychology: Applied, 19(3), 254.

Medina-Borja, A., Pasupathy, K. S., & Triantis, K. (2007). Large-scale data envelopment analysis (DEA) implementation: a strategic performance management approach. Journal of the Operational Research Society, 58(8), 1084–1098.

Medina-Borja, A., & Triantis, K. (2014). Modeling social services performance: a four-stage DEA approach to evaluate fundraising efficiency, capacity building, service quality, and effectiveness in the nonprofit sector. Annals of Operations Research, 221(1), 285–307.

Mumford, E. (2006). The story of socio-technical design: reflections on its successes, failures and potential. Information Systems Journal, 16(4), 317–342. https://doi.org/10.1111/j.1365-2575.2006.00221.x

O'Donnell, C. J., Rao, D. S. P., & Battese, G. E. (2008). Metafrontier frameworks for the study of firm-level efficiencies and technology ratios. Empirical Economics, 34(2), 231–255. https://doi.org/10.1007/s00181-007-0119-4

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

Paradi, J. C., & Schaffnit, C. (2004). Commercial branch performance evaluation and results communication in a Canadian bank—a DEA application. European Journal of Operational Research, 156(3), 719–735. https://doi.org/10.1016/S0377-2217(03)00108-5

Paradi, J. C., & Sherman, H. D. (2014). Seeking Greater Practitioner and Managerial Use of DEA for Benchmarking. Data Envelopment Analysis Journal, 1(1), 29–55.

Probst, T. M., & Brubaker, T. L. (2001). The effects of job insecurity on employee safety outcomes: Cross-sectional and longitudinal explorations. Journal of Occupational Health Psychology, 6(2), 139.

Rasmussen, J. (1997). Risk management in a dynamic society: a modelling problem. Safety Science, 27(2), 183–213. https://doi.org/10.1016/S0925-7535(97)00052-0

Ray, S. C. (1988). Data envelopment analysis, nondiscretionary inputs and efficiency: an alternative interpretation. Socio-Economic Planning Sciences, 22(4), 167–176. https://doi.org/10.1016/0038-0121(88)90003-1

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

Roets, B., & Christiaens, J. (2017). Shift work, fatigue and human error: an empirical analysis of railway traffic control. Journal of Transportation Safety & Security, 0(ja), 1–18. https://doi.org/10.1080/19439962.2017.1376022

Roets, B., Verschelde, M., & Christiaens, J. (2018). Multi-output efficiency and operational safety: An analysis of railway traffic control centre performance. European Journal of Operational Research. https://doi.org/10.1016/j.ejor.2018.04.045

Rosa, R. R. (1995). Extended workshifts and excessive fatigue. Journal of Sleep Research, 4, 51–56. https://doi.org/10.1111/j.1365-2869.1995.tb00227.x

Rubinstein, J. S., Meyer, D. E., & Evans, J. E. (2001). Executive control of cognitive processes in task switching. Journal of Experimental Psychology: Human Perception and Performance, 27(4), 763.

Seaver, B. L., & Triantis, K. P. (1992). A fuzzy clustering approach used in evaluating technical efficiency measures in manufacturing. Journal of Productivity Analysis, 3(4), 337–363.

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

Simar, L., & Wilson, P. W. (2011). Two-stage DEA: Caveat Emptor. Journal of Productivity Analysis, 36(2), 205. https://doi.org/10.1007/s11123-011-0230-6

Sussman, D., & Coplen, M. (2000). Fatigue and alertness in the United States railroad industry part I: the nature of the problem. Transportation Research Part F: Traffic Psychology and Behaviour, 3(4), 211–220. https://doi.org/10.1016/S1369-8478(01)00005-5

Topcu, T. G., & Mesmer, B. L. (2018). Incorporating end-user models and associated uncertainties to investigate multiple stakeholder preferences in system design. Research in Engineering Design, 29(3), 411–431. https://doi.org/10.1007/s00163-017-0276-1

Triantis, K. (2015). Engineering Design and Efficiency Measurement: Issues and Future Research Opportunities. Data Envelopment Analysis Journal, 1(2), 81–112.

Triantis, Konstantinos, Sarayia, D., & Seaver, B. (2010). Using Multivariate Methods to Incorporate Environmental Variables for Local and Global Efficiency Performance Analysis. INFOR: Information Systems and Operational Research, 48(1), 39–52.

Van Dongen, H. P., Maislin, G., Mullington, J. M., & Dinges, D. F. (2003). The cumulative cost of additional wakefulness: dose-response effects on neurobehavioral functions and sleep physiology from chronic sleep restriction and total sleep deprivation. Sleep, 26(2), 117–126.

Wallace, L., Keil, M., & Rai, A. (2004). How software project risk affects project performance: An investigation of the dimensions of risk and an exploratory model. Decision Sciences, 35(2), 289–321.

Wilson, J. R. (2000). Fundamentals of ergonomics in theory and practice. Applied Ergonomics, 31(6), 557–567. https://doi.org/10.1016/S0003-6870(00)00034-X

Wong, M. A., & Lane, T. (1983). A kth Nearest Neighbour Clustering Procedure. Journal of the Royal Statistical Society. Series B (Methodological), 45(3), 362–368.

Yu, M.-M., & Lin, E. T. J. (2008). Efficiency and effectiveness in railway performance using a multi-activity network DEA model. Omega, 36(6), 1005–1017.

Zhao, Y., Triantis, K., Murray-Tuite, P., & Edara, P. (2011). Performance measurement of a transportation network with a downtown space reservation system: A network-DEA approach. Transportation Research Part E: Logistics and Transportation Review, 47(6), 1140–1159.

Chapter 3. Estimating Workload Distribution and Stakeholder Preferences in Autonomous Socio-Technical Infrastructure Systems

Taylan G. Topcu, Ning-Yuan Liu, Konstantinos Triantis, and Bart Roets

Abstract

Today's infrastructure systems are complex sociotechnical systems (STSs) that progressively depend on the cooperation between humans and autonomous systems for their real-time management. Traditionally in STSs, operational decisions are made in real-time by "Controllers" while the long-term production technology changing decisions are made by higher level decision makers. This could potentially introduce strategic misalignments between the real-time and long-term decisions and therefore increase failure risks. The increasing role of automation in STSs is a leading example of this case with a crucial impact on safety. Advances in automation technology, increasing financial pressures, and the desired reduction in human errors incentivize the managers of STSs to increase the role of automation. However, in many instances, dynamic contextual characteristics of the infrastructure network, e.g., traffic complexity, render the use of automation ineffective and require manual Controller interventions to sustain reliable operations. In this paper, we investigate the workload distribution between collaborating humans and autonomous decision-making units. Additionally, we study the influence of contextual variables on Controller preferences with respect to Controllers collaborating with autonomous systems. To address these goals, we propose a novel two-stage framework that combines Data Envelopment Analysis (DEA) with Machine Learning (ML) techniques. Our framework allows us to explore the potential of ML to explain the influence of contextual variables. We apply our methodology to a large-scale dataset from INFRABEL (the Belgian National Railway Company) and verify our results with domain experts to understand Controller preferences regarding the use of automation in safety-critical environments.

Keywords: Data envelopment analysis (DEA), machine learning, autonomous sociotechnical systems, revealed stakeholder preferences.

3.1 Introduction

Infrastructure systems are complex sociotechnical systems (STSs) that rely on the successful collaboration among human decision makers and autonomous systems to deliver critical services (O'Sullivan and Sheffrin 2007). Given today's highly competitive economy, the management and operation of STSs occasionally face conflicting performance pressures of safety, economic efficiency, and quality of work-life. At the same time they encounter uncontrollable environmental phenomena, technological advances, varying operator skills, and societal demands (Oster 1999).

In STSs, safety critical decision-making activities are allocated to specifically trained people denoted as Controllers. While Controllers make daily operational decisions regarding service delivery, STSs are typically managed by large organizations. Long term production technology changing decisions (e.g., increasing automation) are usually made by higher level managers. However, Controllers and managers are subjective individuals therefore might have conflicting or misaligned preferences (Keeney and Raiffa 1993) with those of the organization or with each other. Consequently, individual value maximizing tradeoffs (e.g., cost minimization vs. safety maximization vs. fatigue minimization) of these internal stakeholders could unintentionally lead to the underutilization of capital investments and negate re-organization efforts. A highly relevant example of the described phenomenon is the use of automated decision-making systems to assist in safety-critical human decision making. The increase in automation is a sociotechnical change that is rapidly spreading to the rest of the World's workforce (Frey and Osborne 2017; Acemoglu and Restrepo 2017). Modern STSs, such as real-time railroad infrastructure management systems, are currently in the process of cultivating this change as they increasingly rely on collaborative autonomous systems to handle some of the Controller workload. Therefore, as part of this research, we use these types of STSs as test-beds to investigate and learn about the human-autonomous system collaboration in the management of safety critical tasks.

The primary use of automation in the control of STSs is to decrease system risk levels by handling some of the Controller task workload, which is one of the leading sources of

STS failures (Rasmussen 1997). It is argued that a "safe operation envelope" delineates the accident free operational boundaries of a STSs by three "frontiers": (i) the Controller workload, ii) the economic sustainability, and (iii) the designed safety limit of the artifact or system. Rasmussen's "safe operation envelope" proposes that allocating a safety-critical decision maker with more workload than s/he could handle considerably increases the failure risk of STSs. Therefore, the measurement of the Controller workload is imperative. However, workload measurement in STSs is a challenging task due its complex nature (Leveson 2011). Recently an approach was proposed to quantify the relative Controller task workload based on operational data (Topcu, Triantis, and Roets 2019) by using an empirical productive efficiency measure (Farrell 1957; Charnes, Cooper, and Rhodes 1978) on the Debreu-Farrell (Debreu 1951; Koopmans 1951) frontier. In this paper, we leverage this quantification approach to investigate how the workload is shared between collaborating humans and autonomous Decision-Making Units (DMUs) in modern STSs.

We specifically pursue two research questions. The first is "How does the workload distribution between collaborating human and autonomous decision-making systems vary given dynamic operational demands? We address this question by considering Controllers and the automated systems they collaborate with as parallel transformation process networks. We collect rich and high fidelity operational dataset from Traffic Control Centers (TCCs) of the Belgian national railroad company (INFRABEL). We then employ a Data Envelopment Analysis (DEA) approach to compute how much workload is handled by the people and the autonomous systems, represented by the efficiency scores (Topcu, Triantis, and Roets 2019). The second question is: "What are the revealed Controller preferences regarding the workload delegated to automation, given observed contextual infrastructure network characteristics?" We address this question by establishing the interface between DEA and Machine Learning algorithms (ML) through a novel two stage approach that uses the workload measurement in the first stage. In the second stage of our approach, we relate the task workload to the contextual variables that influence Controller preferences regarding collaboration. Our analytical approach is driven by the unique transformation process we investigate, and which can be summarized with the following.

Each INFRABEL Traffic Controller (TC) manages a workstation that controls a physical portion of the railroad infrastructure and each workstation has a built-in automated decision aid system (ADAS)[9]. The TC can activate the ADAS to perform infrastructure control decisions for a time-period, a train, or a certain node within the controlled railroad network. Given this operational flexibility, INFRABEL instructs TCs to utilize the ADAS to the extent that they feel comfortable with. Since there is no mandatory operation instruction for the TCs, the context, conditions, and magnitude of the workload they choose to allocate to the ADAS vary drastically depending on contextual operational conditions at the time (e.g., Controller fatigue, traffic density, train delays, etc.). To elaborate, the contextual conditions are observed by the TC who then decides based on his/her preferences, which sections of the network needs to be handled manually and which can be delegated to the ADAS. Moreover, each individual TC's ADAS utilization preferences could be different from the preferences of the managers that made the investment decision to install it. Consequently, individual controller decisions can lead to the underutilization of the ADAS and potentially negate the organizational efforts to increase automation.

We believe the combination of factors described above renders this paper an innovative application of DEA based on the following criteria:

The application domain: as for the first time in the DEA literature we study collaborating humans and autonomous decision making systems that are making safety critical infrastructure control decisions. To achieve this, we are opening the DEA black-box of our previous study (Topcu, Triantis, and Roets 2019) where we did not differentiate between the tasks handled by the TC and the ADAS. Our high fidelity and complex application domain is in line with previous efforts to spread the use of DEA (Paradi & Sherman, 2014; Triantis, 2015).

Innovation and the thoroughness of the methodology: The unique transformation process we explore renders this study innovative with respect to many aspects. The first is our treatment of the contextual (Z) variables. Instead of adopting the traditional approach

[9] Also known as "Automatic Route Setting" (Pachl 2002). It sets the train's route when a train approaches a signal.

of treating the Z variables as efficiency influencing or peer selection determining factors (Banker and Morey 1986; Ruggiero 1996; Daraio and Simar 2005; Simar and Wilson 2007; Triantis, Sarayia, and Seaver 2010; Johnson and Kuosmanen 2012), we consider them as the driver of the distribution of workload between TCs and ADASs. This highlights the second innovative aspect of this paper. Since we quantify the workload with Farrell efficiency scores, we can leverage ML techniques to establish a relationship between the contextual conditions that shape the Controller preferences regarding the collaboration with automation. Our approach for utilizing ML is unique, as the studies in our research domain have so far only considered ML as a tool to handle large datasets as discussed in Section 2. Our methodology demonstrates the power of DEA-ML integration to provide highly personalized, flexible, and useful managerial information.

The uniqueness and completeness of the used data: The highly granular data is tailor-made for this study and it differs considerably from the previous modeling of railway control centers (Roets, Verschelde, and Christiaens 2018) and the workstations (Topcu, Triantis, and Roets 2019). Compared to our previous study (Topcu, Triantis, and Roets 2019), the data source is increased to seven fully-operational INFRABEL TCCs instead of a single pilot. Consequently, the dataset is large for a DEA study and consists of 21,930 observations from a purposefully anonymized month in the year 2018. The number of variables in the data have expanded considerably, as it now includes measurements for the Controller age, phone calls, trains delays, and the amount of time that spent using the recently installed forecast tool.

The validation of the proposed methodology and obtained results: We verify our approach by investigating the traffic control process at INFRABEL, on-site within their TCCs. We provide a brief comparison of our approach with traditional regression based 2-stage DEA models (Ray 1988) and check the validity of our ML based interpretation of the contextual variables based on the feedback from domain experts.

Contribution to modeling approaches: We propose a novel two-stage approach that consists of DEA in the first stage and ML in the second stage. We experimented with ML algorithms to explain the influence of contextual variables given the complex transformation process we study. We use permutation importance techniques (i.e., feature

relevance) to rigorously compute the ML interpretation of the contextual variables and compare them to linear regression approaches that are commonly utilized in other 2-stage DEA models (Hoff 2007; Banker and Natarajan 2008; McDonald 2009; Simar and Wilson 2007). Our results indicate that traditional linear regression-based approaches could potentially misinterpret the influence of contextual variables by a large margin in complex production processes found in STSs. Therefore, we also demonstrate the potential benefits of integrating ML techniques with DEA, as its prediction accuracy far exceeds the regression based methods in this context. This capability allows us to establish an interdisciplinary bridge between the systems science and performance measurement literatures.

Policy insights: Infrastructure management systems will become increasingly autonomous. However, the safety critical decisions are expected to be overseen by humans given the severe consequences of their failures (National Transportation Safety Board 2016; Salmon, Walker, and Stanton 2016). From a system safety perspective, our study provides insights from an operational STS regarding the human-autonomous system collaboration in safety critical environments. From a system design perspective, the high fidelity insights obtained from this study could be used to inform the design of high value STSs (Topcu and Mesmer 2018). From a managerial perspective, increasing digitization and recent developments in measurement technology enables enterprises to collect highly disaggregate, precise, and large datasets that capture the actions of lower level decision makers. Such granular data allows one to learn about the revealed preferences of lower level decision makers without the need for self-reporting tools such as surveys. From a data analytics perspective, our use of ML techniques (Breiman 2001a; 2017; Chen and Guestrin 2016a; Vapnik and Lerner 1963; Vapnik, Golowich, and Smola 1997; Nørgård et al. 2000) demonstrates how reliable and useful information (e.g., revealed employee preferences) could be extracted from the operational data to make well informed decisions. Finally, we respond to the calls from the macro ergonomics community to provide holistic systems driven approaches to address the interdisciplinary nature of STSs (Leveson 2011; Kleiner et al. 2015).

The rest of the paper is organized as follows. Section 2 provides a concise literature review. Section 3 describes the transformation process and the employed methodology along with the description of the data. Section 4 presents the results while Section 5 concludes.

3.2 Literature Review

Controllers make safety-critical and service-oriented decisions for STSs that ensure the delivery of system level services. TCs at INFRABEL, like many other STS Controllers, handle their workload by collaborating with an automated system. By definition, safety critical decisions are context dependent and usually unique. Thus in STSs, this role is allocated to a group of humans (Wilson 2000). This couples the performance of the system with the performance of the Controllers that are subject to work environment related issues. Human performance is influenced by social factors such as fatigue (Ferguson et al. 2008; Roets and Christiaens 2017), mental stress (Beehr 2014), situational awareness (Salmon et al. 2009), prospective memory (Grundgeiger, Sanderson, and Dismukes 2015), among many others. In short, the TC workload emerges from the interaction of the social network and the technology (Kroes et al. 2006) and it is consequently of an interdisciplinary nature (Leveson 2011).

In this research, we use DEA (Charnes, Cooper, and Rhodes 1978) that is rooted in microeconomic production theory (Koopmans 1951; Debreu 1951; Farrell 1957) to understand the TC workload. We consider the work handled by the TCs as an abstract transformation process that converts a set of resources (inputs) into a set of outputs/services/outcomes under uncontrollable work environmental factors that are described in detail elsewhere (Topcu, Triantis, and Roets 2019). Thus, we start our literature review with performance measurement approaches that consider contextual heterogeneity.

3.2.1 DEA Methods that Deal with Environmental Heterogeneity

The idea of considering differences in the performance environments was initially proposed through a single categorical variable that restricts the reference set of the evaluated DMU (Banker and Morey 1986). In the following years, researchers proposed

three main approaches among others to consider efficiency performance measurement under environmental heterogeneity. The first is the 2-stage model (Simar and Wilson 2007; 2011). The 2-stage approaches compute the efficiency on the first stage without consideration of the Z variables, and then investigate the influence of contextual variables through linear regression methods. This second stage is predominantly based on truncated regressions (Simar and Wilson 2007, 2011), but also applies Ordinary Least Squares (OLS) and Tobit regressions (McDonald 2009; Banker, Natarajan, and Zhang 2019). The second is the multivariate method (Seaver and Triantis 1992; Triantis, Sarayia, and Seaver 2010). In the multivariate approach, the comparability assumption is rigorously tested through robust statistics (Hubert, Rousseeuw, and Branden 2005). The DMUs in the production possibility set are then grouped into relatively homogenous subgroups through clustering methods (Herrera-Restrepo et al. 2016). The efficiency scores are analyzed both within these clusters and with respect to the meta-frontier that disregards the differences in performance environments (O'Donnell, Rao, and Battese 2008). The third and final approach is the semiparametric method (also known as the one stage method) (Johnson and Kuosmanen 2011; 2012). The primary advantage of the one stage method is the computation of efficiency scores in conjunction with the influence of the Z variables.

While this paper also deals with contextual variables, it is fundamentally different from the literature discussed above. First, driven by the specific transformation process, this paper is concerned with understanding the influence of contextual variables on the workload distribution decisions of a TC among the ADAS and him/herself. We represent the collaborating TC and the ADAS as a mutually exclusive parallel process represented by two nodes. Second, this paper differs in terms of the analytical method used in the second stage. The traditional two-stage approaches a usually rely on some variation of a linear regression approach in the second stage to explain the efficiency scores that were calculated in the first stage. While this approach is widely adopted by relatively simpler transformation processes such as banking and management, we question the applicability of this approach to STSs given their complex, nonlinear, and automated operational characteristics. Similar points have been raised by previous research on transportation systems (Karlaftis and Vlahogianni 2011) and safety-critical infrastructures (Paltrinieri, Comfort, and Reniers 2019). In contrast to linear regression-based approaches, the ML

approaches explored in this study are able to handle complex, non-linear relationships between the efficiency scores and Z variables. Similar to DEA, our approach lets "the data speak for itself", and does not assume a priori a linear relationship between efficiency scores and Z variables.

3.2.2 DEA and Machine Learning

Machine learning (ML) is a term used generally for automated statistical predictive modeling. A review of fundamentals (Alpaydin 2009) and an extensive literature review is provided elsewhere (Kotsiantis, Zaharakis, and Pintelas 2007; Tan 2018). ML has three types of learning: supervised, unsupervised, and reinforced. Supervised ML establishes a predictive statistical relationship between predictor variables and target variables[10]. Examples include traditional statistical regression algorithms and perception-based algorithms (e.g. artificial neural networks). Unsupervised learning does not seek a target variable to predict, instead, it is used for segmentation/sub-setting purposes. Widely used clustering techniques such as k-means (Hartigan and Wong 1979) belong in this group. The final type of ML is reinforcement learning. This is similar to artificial intelligence, as the algorithm is designed to capture the best possible knowledge by accumulating experiences and learning from it by trial and error. While ML already made a significant impact in a wide array of application areas, the literature on the intersection of ML and DEA is relatively scarce.

One of the first studies that experimented with ML in DEA investigated the use of clustering techniques to identify reference sets in which the DMU homogeneity assumption holds (Seaver and Triantis 1992; Triantis, Sarayia, and Seaver 2010). This method was recently implemented and verified on a fully operational infrastructure management system (Topcu, Triantis, and Roets 2019). A considerable group of publications explored the complementary and contrasting roles of the neural network algorithms and the DEA regarding the assessment of productive performance (Athanassopoulos and Curram 1996; Costa and Markellos 1997; Santin, Delgado, and Valino 2004). Future performance of DMUs were investigated from an ex-ante perspective through a predictive approach by

[10] Denoted as independent and dependent variables respectively in the statistics jargon.

using a decision tree learning model (Sohn and Moon 2004). The use of ML classification and prediction techniques to identify valuable sources of organizational information was demonstrated for a DEA based decision support tool (Samoilenko and Osei-Bryson 2013). The predictive ability of neural networks was utilized to forecast the future performance of DMUs for improved personnel selection (Azadeh et al. 2011).

We observe that a significant portion of ML-DEA papers are focused on computational speed issues when dealing with large datasets. Two big data extensions of regular DEA approaches were introduced, where both proposed to divide the data to identify observations on the frontier iteratively (Khezrimotlagh et al. 2019; Zhu, Wu, and Song 2018). Similarly, a subset of the data was used to train ML algorithms to predict the efficient observations on larger datasets (Zhu et al. 2018). While we find the approach intriguing, we are concerned that the results might be biased given the selection of the training subset. In a similar line of inquiry, the potential benefits of incorporating ML methods to understand environmental efficiency was discussed by Song et al. (2018).

Given this concise review, we believe that there is an untapped potential regarding the use of ML in efficiency measurement and it could be leveraged to improve both theory and practice. A primary goal of this paper is to explain the influence of contextual variables on the employee preferences in complex transformation processes using supervised ML techniques. In complex DMUs, the correlation between the variables are low and the operational conditions are highly dynamic. Thus, we utilize the predictive power of supervised ML algorithms and use contextual Z variables as predictors of the DEA efficiency scores that represent how much workload is handled by the ADAS and the TC. Our approach is translational, as we conduct this research on a fully operational STSs where humans and autonomous decision-making units collaborate towards a common goal. We experiment with four well established machine learning algorithms: (i) the random forest (Breiman 2001b), (ii) Xgboost (Friedman, Hastie, and Tibshirani 2000; Friedman 2001; Chen and Guestrin 2016), (iii) the support vector machine (Boser, Guyon, and Vapnik 1992; Guyon, Boser, and Vapnik 1993; Cortes and Vapnik 1995; Vapnik, Golowich, and Smola 1997), and (iv) artificial neural networks (Jain, Mao, and Mohiuddin 1996). We leave the discussion of the respective methodologies associated with these

algorithms, why we considered these algorithms appropriate, and our proposed technique for calculating the relative influence of predictor variables in §3.3.

3.2.3 Revealed Stakeholder Preferences

Large STS organizations such as INFRABEL are composed of a vast number of decision-makers with subjective and potentially conflicting preferences (Franssen 2005). Since STSs are rational organizations (Scott 2015), organizational preferences are relatively easy to identify. For INFRABEL, operating the infrastructure system safely and on schedule is an easily identified preference. This is reflected in the company's strategic objectives "safety first" and "trains on time". On the other hand, TCs are shift workers who are subject to work environment issues such as fatigue (Dawson and McCulloch 2005) and traffic complexity. Thus, we expect TC preferences to be dynamic and context dependent. From a managerial perspective, one could capture the TC preferences through surveys or interviews. However, the "stated preferences" acquired from self-reporting methods could significantly differ from actual daily behavior. In this paper, we specifically focus on TC preferences that are associated with collaborating with the ADAS. Instead of administering surveys, we analyze a large operational dataset from INFRABEL in pursuit of learning about "revealed preferences" (Samuelson 1948; Chambers and Echenique 2016) that describe how TCs actually behave given uncontrollable contextual factors represented by the Z variables.

3.3 Methodology

3.3.1 The Data

The highly granular data were custom made for this study and are from seven fully-operational INFRABEL TCCs for an entire month in the year 2018. It includes 21,930 observations that represent one operational hour for both the ADAS and the TC's that make infrastructure management decisions. The data used in this paper differ considerably from the initial modeling of railway control centers (Roets, Verschelde, and Christiaens 2018) and our previous study at the workstation level (Topcu, Triantis, and Roets 2019). We added the train delay, responded phone calls, and the age of the controllers specifically for

the purpose of this research. Table 3-1 provides variable definitions and descriptive statistics.

Table 3-1 The Sociotechnical Data and its Descriptive Statistics

Variable Name	Description	Mean	Range
Controller Skill Days	We measure skill levels with the number days the TC was assigned for the designated role.	518.00	[29;759]
Signal Passes	Number of total signal passes of trains within the control area	44.49	[0;193]
Manual Movement Decisions	Time spent (in seconds) for manually opening signals by TCs	111.04	[0;1580]
Auto Movement Decisions	Time spent (in seconds) for manually opening signals by the ADAS. The time needed for each signal is considered the same as when opening a signal manually.	277.17	[0;1870]
Adaptation Decisions	Time spent (in seconds) for decisions that change the state of the railroad such as merging or splitting trains, re-routing of trains, or special procedures at single-track lines. Performed manually by TCs.	186.57	[0;1354]
Anticipation	Measure (in seconds) of TC time spent using the forecast tool that anticipates the future state of the network.	4.96	[0;840]
Responded Phone Calls	Number of phone calls addressed by TCs. TCs routinely receive phone calls from other INFRABEL personnel about decisions that require further information.	1.43	[0;27]
Age	Age of the Controller	47.03	[23;63]
Traffic Complexity	Measure of traffic complexity of control area. Estimated by using the number of control signal passes and performed adaptation decisions	926.06	[0;72000]
Traffic Density	Measure of traffic density in control area. Calculated by dividing the number of train movements with the number of large traffic control signals controlled by that Controller.	832.34	[0;10605]
Fatigue Level	This variable represents the mental fatigue of TCs. It is calculated by INFRABEL's predictive tool that is conceptually based on the fatigue Risk Index (Roets and Christiaens 2017; Folkard, Robertson, and Spencer 2007).	0.85	[0.64;1.55]
Delay	Average train delays within the control area. Measured from the scheduled time in seconds. Large delays are due to freight trains. Trains running before schedule (negative delays) can also perturb traffic flows, and therefore are considered through their absolute value.	562.17	[0;25,936]

3.3.2 The Analytical Approach

Going back to our first research question, this paper investigates how the total workload is distributed between the TCs and the ADAS, over time. We model the traffic control process of the TCs for two parallel perspectives, the workload the TC chooses to handle manually and the workload delegated to ADAS by the TC. Our analytical approach is composed of two steps. We present the two steps in Figure 3-1. In the first step, represented with the blue box, we frame two parallel and mutually exclusive DEA models that represent the ADAS and the TC. Notice that the blue rectangle frame in Figure 3-1 is exactly the

same with the frame of the workstation boundary in our previous work. However, the realism of the model is increased with additional variables (Topcu, et al. 2019). The temporal measurement unit of the DMU is selected as one hour because: (i) it is narrow enough to capture the dynamic changes in the state of the infrastructure and (ii) it is large enough to capture the entire collaborative control process. The horizontal arrows on the left side represent the resources used by each process (X variables). The only irrevocable resource used by the manual process is one hour of TC expertise. For the ADAS model, the only input variable is the total signal passes of trains within the control area. Arrows on the right side of black boxes represent the outputs (Y variables) generated by each process. There are four main tasks for the TCs: a) making "movement decisions", which is identical with those performed by the ADAS; b) making track "adaptation decisions", which can only be performed manually; c) "forecasting/anticipating" the future state of the network through a recently installed projection tool; and d) "responding to phone calls" from other personnel to provide information. The ADAS can only perform movement decisions in its current version.

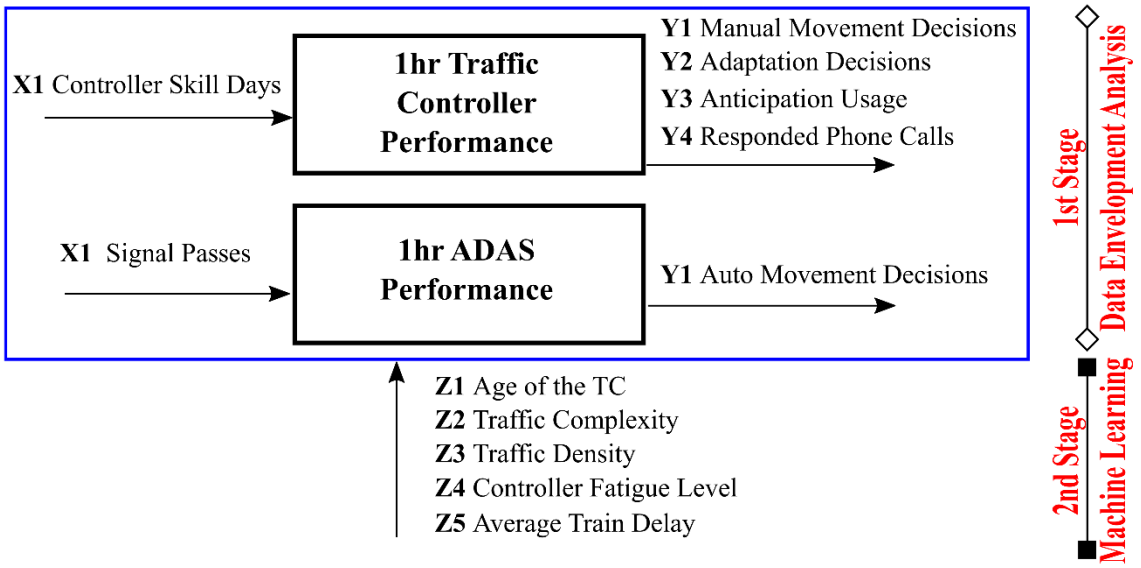

Figure 3-1 Two-stage Methodology and Sociotechnical Variables

While there are various models in the DEA literature to measure performance in the presence of contextual variables (§2.1), we purposefully disregard them in the first stage of our model. This paper explores additional uses for the contextual variables in DEA to

obtain information the managers can use to make rational decisions (e.g., tendencies of their employees (TCs) to work with the ADAS). Therefore, guided by the characteristics of the collaborative transformation process, instead of investigating the influence of contextual variables on efficiency scores we consider them from a TC preference revealing point of view. We employ an output-oriented Variable Returns to Scale model (Banker, Charnes, and Cooper 1984) to compute the efficiency scores based on the assumption that experienced TCs should be able to handle more workload compared to inexperienced TCs. We select the type and orientation of our ADAS model based on a similar assumption. The ADAS should perform increasingly more movement decisions given that one provides increased traffic volume to it. Thus, for both the TC and the ADAS we formulate the linear optimization formulation as:

$$\text{Max } \theta \tag{1}$$

$$\text{Subject to } \sum_{j=1}^{n} x_{ij} * \lambda_j \leq \mathrm{x_{ij_0}} \ \forall i = 1 \ to \ m \tag{2}$$

$$\sum_{j=1}^{n} y_{rj} * \lambda_j \geq \theta * y_{rj_0} \ , \forall r = 1 \ to \ s \tag{3}$$

$$\sum_{j=1}^{n} \lambda_j = 1 \tag{4}$$

$$\lambda_j \geq 0 \ , \theta \geq 0 \tag{5}$$

In equations 1-5; n represents the number of DMUs, x_{ij} represents the amount of resource i consumed by the j^{th} DMU, λ_j is the weight assigned to a peer j, θ is the efficiency score, $y_{rj}$is the r^{th} output of j^{th} DMU, and y_{rj_0} is the r^{th} output of the DMU under investigation.

Recalling our second research question, we are interested in capturing TC revealed preferences regarding the use of automation based on the state of their performance environment represented by Z variables (shown as vertical arrows in the bottom half of Figure 3-1). Since each TC performs traffic management decisions on their dedicated portion of the network and have a dedicated ADAS at their disposal, the extent that each TC chooses to utilize the ADAS depends strictly on their preferences. While we realize the possibility of investigating TC and ADAS workloads concurrently, which could require multi-output type regression (Borchani et al. 2015), we consider this paper as a naïve first step and treat TC and ADAS workloads as mutually exclusive. We proceed to describe the experimented learning algorithms and their validation.

3.3.3 Selection and Validation of ML Models to Extract Revealed Preferences

In the second stage of our analytical approach, we investigate the TC preferences regarding the workload distribution between the ADAS (θ_{ADAS}) and the TC (θ_{TC}). Following the ML terminology for supervised learning, we treat the DEA contextual Z variables as predictor variables and formulate two mutually exclusive predictive models that consider the efficiency scores of the TC (θ_{TC}) and the ADAS (θ_{ADAS}) as target variables. The methodology of the second stage could be summarized with the following steps: (i) implementation of the baseline ML algorithms, (ii) tuning of the ML algorithm and its validation, (iii) identification of the importance of features (variables) for each specific algorithm. An overview of ML model calibration and how it fits with the proposed DEA approach is provided in Figure 3-2. Figure 3-2 depicts how we treat the contextual variables in predicting the efficiency scores calculated in Figure 3-1, using the algorithms described in this subsection.

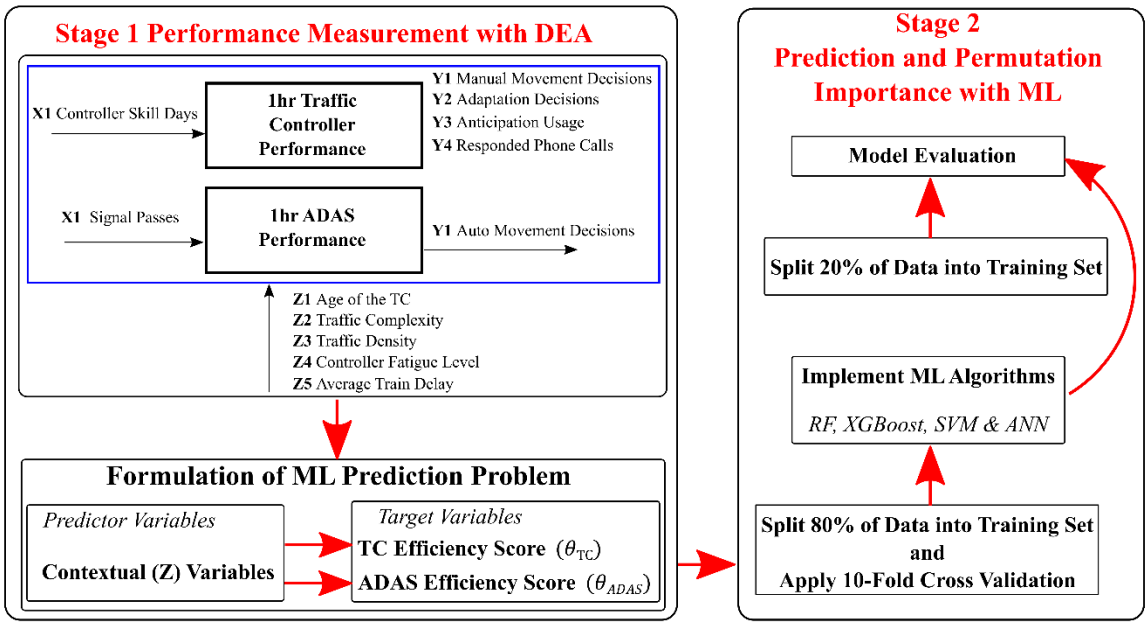

Figure 3-2 An Overview of Machine Learning Algorithm Calibration

Before covering the details of the implemented algorithms, we start with our algorithm validation approach. In general, ML algorithms divide the data into two sub-sets, i.e., the training set and the test set. The algorithm is then allowed to learn from the training set and its parameters are optimized before exploring its prediction accuracy on the test set. For all

implemented algorithms in this paper, we randomly split 80% of the data for the training set and the rest of the data for the test set. We first employ a random search strategy for the hyperparameter[11] tuning since this approach has been proved to be more effective than the traditional grid search (Bergstra and Bengio 2012). Second, we apply a K-fold cross-validation (Kohavi 1995) to obtain an unbiased evaluation of the model performance and to prevent overfitting issues. The process can be summarized with the following. We divide the training set into 10-folds, meaning that 90% of the data are treated as the actual training set and the remaining 10% as the validation set. The machine learning model is then iteratively trained and validated on these ten different folds. Lastly, for all implemented algorithms, we test the bias and accuracy of the test set based on well-known statistical considerations shown in Equation 6, i.e., the mean absolute error (MAE) and the root mean squared of error (RMSE). In Equation 6, $\hat{y}_i$ is the predicted value, y_i is the actual value and n is the number of observations. The values of these two factors indicate the prediction accuracy of the implemented model within the training and test datasets.

$$MAE = \frac{\sum_1^n |y_i - \hat{y}_i|}{n} \; ; RMSE = \sqrt{\sum_1^n \frac{(y_i - \hat{y}_i)^2}{n}} \tag{6}$$

The next step is the extraction of "feature importance" once the ML models are optimized and validated. Feature importance in the ML language represents the impact of predictor variables on the target variables which, correspond to the Z variables and the efficiency scores in our case. While there are various methods to compute the feature importance such as the variance and Gini importance (Breiman 2001b), most approaches are prone to bias (Strobl et al. 2007). We use the permutation importance concept because it normalizes the biased measure through a permutation test that iteratively replaces a predictor variable with a random value, repeats the analysis, and returns significance P-values for each predictor variable (Altmann et al. 2010). Thus, we select the permutation importance test because it is relatively less susceptible to bias compared to other alternatives. Permutation test captures the importance of a predictor variable by calculating

[11] Hyperparameter is a parameter whose value is set before the learning process begins. In other words, by setting up a grid of hyperparameter values and selecting random combinations to train, we can minimize the estimation error rates.

the overall goodness of fit (R^2) following Equation 7 and then calculating the difference in the fit between the baseline value and the randomized value. Hence, the computed relative permutation importance for each predictor variable is always positive, does not sum up to one, and finally, indicates the relative predictive strengths of each variable. We next cover the implemented algorithms.

$$R^2 = 1 - \frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2} \quad (7)$$

The first is the Random Forest (RF) and in this paragraph we describe why we considered it as a promising algorithm to experiment with. The RF algorithm belongs to the family of ensemble methods, which rely on the aggregation of individual decision trees (Breiman 2001b). We define the individual decision trees used in this paper, following a classification and regression tree model of form $f(x; \Theta_k)$. The details of the decision trees are described elsewhere (Breiman 2017). The RF algorithm is composed of a collection of random trees, where each unpruned tree is assigned a subset of features and a sub-dataset. Within this random forest, predictions of all individual trees are aggregated through majority voting to determine the collective result of the forest. Thus, instead of being influenced by faulty prediction of a single decision tree, RF demonstrates good stability on high-dimensional datasets, avoids overfitting issues, and yields a high accuracy rate (Ali et al. 2012). These characteristics also provide the reasons why we consider this algorithm appropriate for this paper. The RF is defined with L collection of decision tree predictors $\{f(x; \Theta_k), \mathrm{k} = 1, \ldots, \mathrm{L}\}$, where Θ_k are independent random vectors, and x represents the predictor variables (that correspond to Z variables in our case). Each tree is randomly generated by selecting a set of features (m) from the set of Z variables in our DEA notation, $m \in Z$ and a random sample of size n from the training dataset $n \in N$, where N represents the total number of DMUs in the training set. The RF prediction is then obtained by averaging K decision trees as shown in Equation 8. We stop the optimization and select the ideal number of decision trees when the expected variance converges.

$$\hat{f} = \frac{1}{K}\sum_{k=1}^{K} f(x; \Theta_k) \quad (8)$$

The second algorithm is the extreme gradient boosting (Xgboost) (Chen and Guestrin 2016b) that is also from the family of ensemble methods. Since its publication, Xgboost

had a strong impact on the ML community due to its computational speed when dealing with large datasets (Babajide Mustapha and Saeed 2016; Sheridan et al. 2016). Xgboost builds on the core idea of the gradient boosting machine algorithm (GBM) (Friedman, Hastie, and Tibshirani 2000; Friedman 2001), but it solves the inefficiencies of GBM by: (i) adding a regularization term to prevent overfitting, (ii) a parallel processing feature to minimize computational time, and (iii) allowing to customize algorithm parameters for increased flexibility. These features render the XGBoost algorithm an appropriate choice for the complex STS investigated in this paper. The objective of the Xgboost algorithm is to calculate the prediction scores and it is provided in Equation 9. In Equation 9, the first term on the right-hand side represents the loss function (l) that measures the difference between the predicted value $\hat{y}_i$ and the target value y_i. The second term is composed of the regularization term (Ω) and the prediction of a single decision tree (f_k). The role of the regularization term is to penalize the complexity of the model to avoid overfitting and make the prediction stable.

$$Obj(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \quad (9)$$

The regularization term (Ω) in Equation 9 is defined with Equation 10. It is controlled by the characteristics of the decision trees that are: the constant (γ), the number of leaves (T), the score on each leaf (ω), and the degree of regularization (λ). In addition to these parameters, we also use the hyperparameters of shrinkage and column subsampling, which we optimize through the random search to prevent overfitting issues. Shrinkage scales added new weights after each step of the tree boosting, to make the model learn slower and better. Column subsampling can accelerate the training process by only considering random subsets of descriptors when building a given tree (Friedman 2002).

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda\|\omega\|^2 \quad (10)$$

The third algorithm we implement is the support vector machine (SVM). While SVM was initially proposed to handle nonlinear generalization problems (Vapnik and Lerner 1963) it was later developed into a supervised learning model for classification and regression with high-dimensional data (Boser, Guyon, and Vapnik 1992; Guyon, Boser, and Vapnik 1993; Cortes and Vapnik 1995; Vapnik, Golowich, and Smola 1997). In SVM, vectors of predictor variables are transformed into hyperplanes that exhibit linearity and

regression is applied within this hyperspace. While the hyperplanes exhibiting linearity assumption of the SVM is contrary to the nonlinear data set we are concerned with, we still choose to include the algorithm in our study for exploration and comparison purposes. The activation function (f) in SVM is defined with Equation 11. In Equation 11, $\phi(x)$ is a nonlinear function that is mapped into a high dimensional space, w is the weight vector, and b is the hyperplane. However, the details are described elsewhere (Cortes and Vapnik 1995). One of the well identified issues with SVM is the necessity of selecting a kernel function beforehand (e.g., radial basis function (RBF), polynomial, linear etc.) and the sensitivity of SVM results to the kernel. The selection of the kernel complicates getting the optimal model configuration in certain instances (Mountrakis, Im, and Ogole 2011).

$$f(w,b) = w \cdot \phi(x) + b \tag{11}$$

The fourth algorithm is artificial neural network (ANN). ANN was initially introduced to explain information processing mathematically that is derived from human biological systems (McCulloch and Pitts 1943; Rosenblatt 1958). Over the years, a variety of ANNs have been formulated for various problems (Jain and Mao 1996). In this paper, we use one of the most utilized ANN algorithms, the multilayer perceptron (MLP) due to its ability to forecast and classify with precision (Nørgård et al. 2000). The basic unit of the MLP algorithm is an artificial neuron (also denoted as nodes). Artificial neurons, like a biological neuron, interconnects to nodes in other layers and transfers the information unidirectionally. MLP algorithms consist of at least three layers of nodes. An illustration of the model for the variables in this paper is depicted in Figure 3-3.

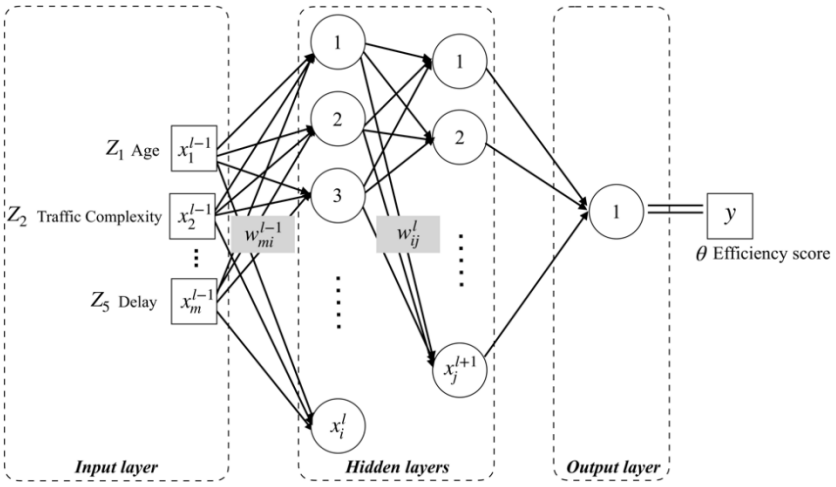

Figure 3-3 Layers of the MLP Algorithm

In the standard MLP model (Hornik, Stinchcombe, and White 1989), the input of neuron j in layer $l+1$ is represented with Equation 12. In Equation 12, w_{ij}^{l} represents the weight of the connecting neuron i in layer l to neuron j in layer $l+1$, and w_{bj}^{l} is the bias of neuron j in layer l, and f is an activation function that represents the nonlinear mapping between the input vector and the output vector. Based on this standard model, we employ the "back-propagation" technique (Gardner and Dorling 1998) that minimizes the entire neural networks' output error. For each training iteration, the algorithm makes a prediction, measures the error for that step, and then goes through each layer in reverse to measure the contribution from each node. We experiment with the ANN based on its ability to detect complex nonlinear relationship between predictor and target variables (Tu 1996).

$$x_j^{l+1} = f(\sum_i w_{ij}^{l}\, x_i^{l} + w_{bj}^{l}) \tag{12}$$

Finally, we implement a simple multi-linear regression (MLR) model using the contextual variables and the efficiency scores. Our purpose is twofold: (i) to demonstrate the inability of linear models to understand the inherently complex nature of STSs and (ii) to provide the audience with a measure that we can use to compare to the ML results that is commonly utilized with the 2-stage DEA approaches. The MLR model is defined with Equation 13. In Equation 13, n represents the data point, Y_n represents the efficiency scores obtained from stage 1, x_{ni} are the contextual variables, β_i are the coefficients for each x_i and $\varepsilon_n$are the random error with expectation 0 and variance σ^2.

$$Y_n = \beta_0 + \sum_{i=1}^{5} \beta_i x_{ni} + \varepsilon_n \quad (13)$$

3.4 Results and Discussion

3.4.1 Comparison of Efficiency Scores - The Workload Distribution

There are 21,930 observations in our dataset that correspond to one hour of traffic control activities. We calculate the two mutually exclusive productive efficiency scores that represent task workloads for the TC and the ADAS, following the procedure described in §3.2. Strictly speaking, the Farrell efficiency scores should be interpreted as the relative distance to the boundary of attainable workload, and therefore varies between 0 and 1 (see also § 3.2). This rescaling allows for a direct and sensible comparison between the respective workloads of the manual and automatic processes. The distribution of the TC workload is observed to be slightly skewed to the right; it has a mean of 0.19 and a standard deviation of 0.16. Low average workload scores are attributed to the highly disaggregate measurement window employed in this study and unavoidable periods where the TC permanently monitors the automated process and remains standby to intervene manually. The average workload for the ADAS is 0.37 that is almost twice the average of TC workload. The median workload for the ADAS is almost identical with its mean indicating a symmetrical distribution with a standard deviation of 0.18. The correlation between two workload measurements is equal to -0.06. We consider this as an interesting result. Our initial expectation was to find a negative correlation between workload distributions based on our empirical observations of the workplace environment. This result indicates that the workload variation is more intricate than what we intuitively predicted. Our expectation was that, at any point in time, increases in the manual TC workload would be linked to decreases in the ADAS workload and vice-versa. However, this was not the case. As the distribution of workload might be influenced by time-of-the-day effects, we provide the boxplots for both TC and ADAS workload over time for a single anonymized workstation in Figure 3-4.

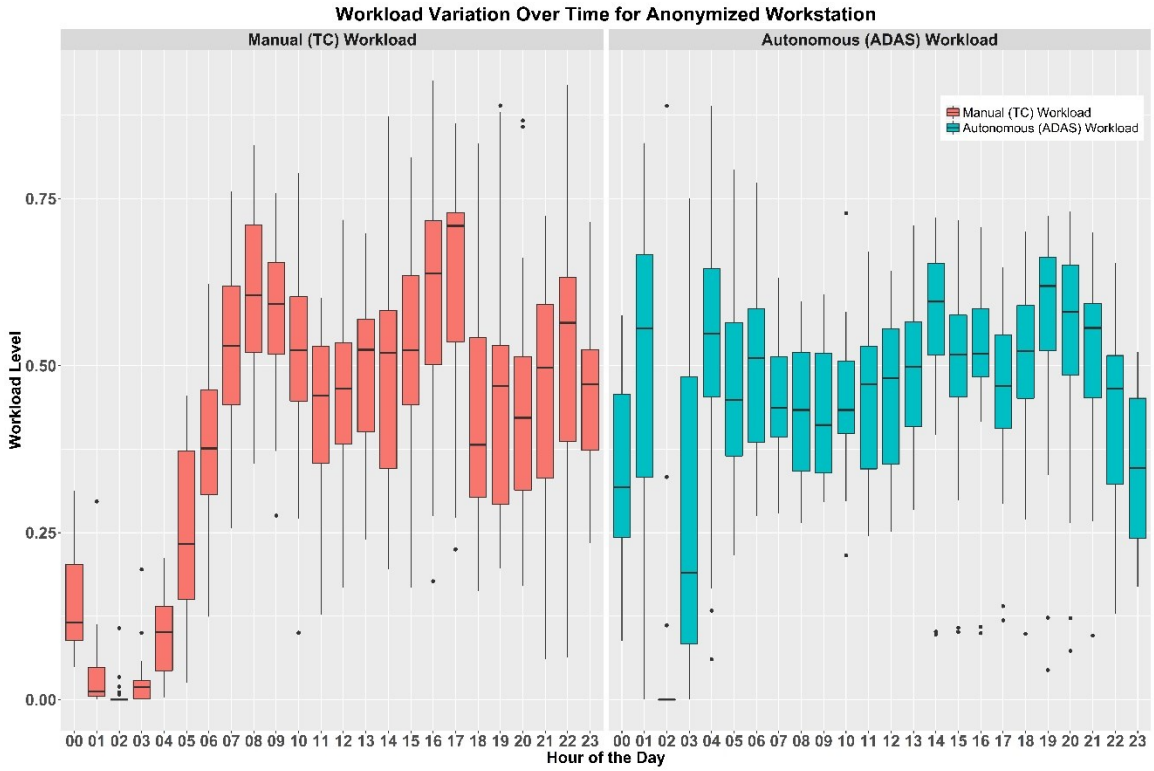

Figure 3-4 Workload Distribution over Time for an Anonymized Workstation

The red box-plots in Figure 3-4 indicate low manual TC workloads around idle traffic hours (1-3AM) and how a steep climb starts after 5 AM, the morning rush hour. We observe two peaks around 8AM and 5 PM that correspond to the rush hour traffic and on average, the TC workload stays relatively high until 6 PM. On the other hand, the ADAS workload represented with turquoise boxes, exhibit high variation during the idle hours (1-3AM) then stays relatively flat until the PM rush hour traffic. Interestingly, the ADAS workload peaks around 7 PM, after the rush hour traffic during which the manual workload stays relatively low, parallel to our intuitive expectations. We discuss some potential sources of this trend in the next subsection. Notice that data for multiple TCs operating the same workstation were used to populate Figure 3-4, thus it could be masking certain individual TC attitudes due to aggregation effects. However, our granular analytical approach provides a plethora of information that allows to focus on the tendencies of individuals, workstations, shifts, or TCCs for a desired period of time. Providing a rigorous, useful, and flexible tool for managers.

From a practical and implementation point of view, Figure 3-4 demonstrates the power of our "parallel DEA model" in terms of quantifying not only the Controller manual workload but also the Controller use of automation, while visualizing the corresponding temporal and spatial variations. This provides valuable operational insights for the STS management, and enables them to identify and implement opportunities for improvement. For example, the temporal fluctuations for the workstation in Figure 3-4 can be compared with the other workstations in the control center, and as such allow for a more optimal work reallocation for the infrastructure network over these workstations. By spreading the (manual and automated) workload more evenly over the entire TCC, Controllers will experience a more balanced work environment, which could lead to improved employee satisfaction and well-being. Our approach also reveals the workstations and corresponding infrastructure where automation is underused (depending on the time of day). This spatial and temporal pinpointing of problem areas can support management in setting up improvement strategies (e.g., increase staff training where necessary, or investigate and eliminate shortcomings in the automation system).

3.4.2 Performance of ML Algorithms in Explaining Revealed Controller Preferences

The second research question this paper investigates is how to capture and explain revealed stakeholder preferences regarding the use of automation given the contextual state of the infrastructure network they control. In other words, we are interested in understanding to what extent contextual variables, that are uncontrollable for the TC yet describe the conditions in which performance takes place, influence the daily operational decisions in terms of the delegation of work between the ADAS and the manual TC. We start our discussion with the validation of ML models. We implement all discussed ML algorithms in Python using the "Sklearn" package (Pedregosa et al. 2011).

For the RF algorithm, the number of trees and the number of features used in each node are the leading parameters. There is a positive correlation between the number of trees and the predictive performance of the model. However, this comes with a computational cost. Thus, we optimize the number of trees for model calibration. In Figure 3-5, we provide the learning curve plot of the RF model that demonstrates the relationship between the

expected variance vs. the number of trees in the forest. Figure 3-5 demonstrates that the expected variance saturates around 0.65 when the number of trees exceeds 40. Thus, we consider this number of trees sufficient and terminate the algorithm. Our calibration efforts through random search yields the optimum hyperparameters that shape the manual TC interventions and ADAS model respectively as: the number of trees [100 (TC), 900 (ADAS)], the max depth [70 (TC), 110 (ADAS)], the max features to two for both, the minimum samples of leaf [3 (TC), 2 (ADAS)], and the minimum samples of a split [2 (TC), 3 (ADAS)]. The difference in the numbers of trees between the manual TC interventions and the ADAS shows that ADAS needs more trees to reach convergence, which indicates the consideration of non-linearities. In other words, we attribute the large number of trees and the depth to the nonlinear and complex nature of the data, especially in the case of the ADAS usage (Oshiro, Perez, and Baranauskas 2012).

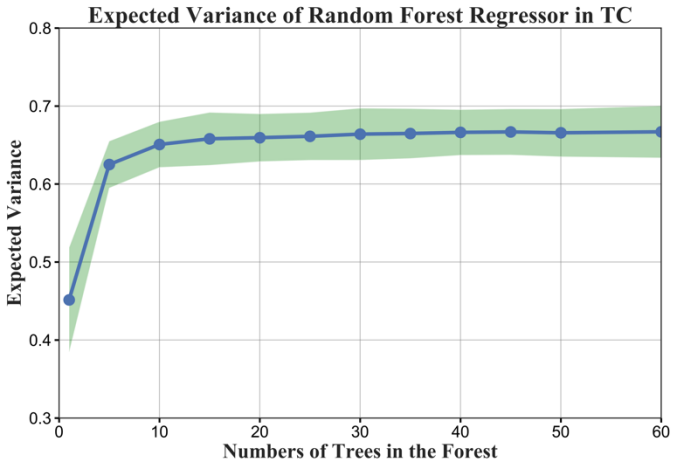

Figure 3-5 Expected Variance vs. Number of Trees in the Random Forest Algorithm

For the Xgboost algorithm, a similar process as RF is implemented for tuning the parameters for both DEA models respectively. The following results of the training model were obtained: the number of trees [600 (TC), 3000 (ADAS)], the max depth [4 (TC), 5 (ADAS)], the shrinkage [0.3 (TC), 0.1 (ADAS)], the gamma [0.05 (TC), 0 (ADAS)], and the subsample [0.8 (TC), 0.9 (ADAS)]. We observe that the ADAS requires a larger number of decision trees with an increased maximum depth than the manual TC intervention model, indicating that the ADAS data are more complex than the manual TC dataset. This resonates with our on-site observations, as the ADAS utilization trends of

TCs vary considerably. The gamma or (the min split loss) represents the minimum positive loss reduction required to make a split on a node. We observe that gamma takes lower values for the ADAS model, indicating a more conservative behavior from the algorithm compared to the manual TC intervention model. The latter two parameters of shrinkage and subsample prevent overfitting with smaller values.

Next, we implement the MLP algorithm using the sequential model in the "Keras" package of Python (Chollet 2015). We specify two hidden layers with 125 and 25 nodes and use a linear activation function for the output layer of both TC and ADAS MLP models. While we recognize that additional layers could be specified to extend the problem into a deep learning problem, the heuristics indicate that two hidden layers are appropriate to examine complex phenomena (Lippmann 1987) including STSs. We consider this level of elaboration sufficient given computational speed issues. We provide the MAE and RMSE for the manual workload over MLP time steps (epochs) in Figure 3-6. Figure 3-6 demonstrates that the MAE and the MSE decreases with each iteration on the training set. The slope for both error measures for the test set demonstrates acceptable fluctuations, thus we terminate the algorithm after 40 epochs for MAE equal to 0.070 and MSE 0.0095.

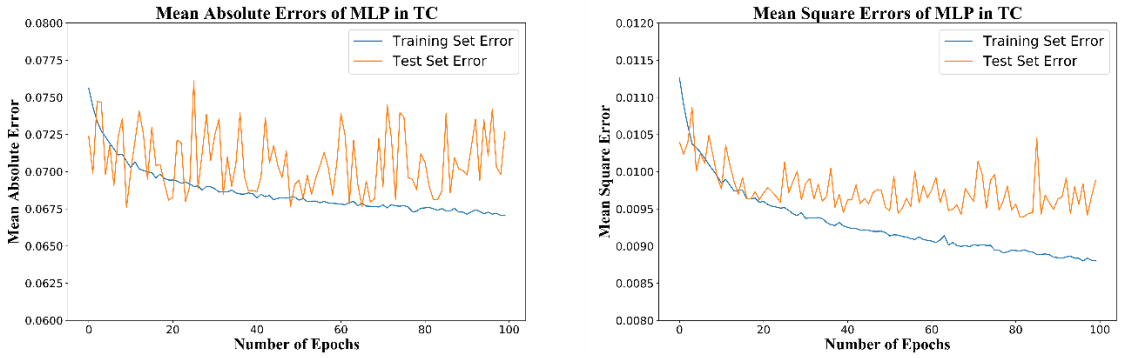

Figure 3-6 Mean Absolute and Mean Square Error for MLP

Finally, the SVM algorithm parameters were optimized following a similar methodology. The leading algorithm parameters are: the constant C, kernel function, and gamma. The constant C controls the smoothness of the activation function that penalizes for the error term. Smaller values of the constant C indicates that the optimizer seeks larger-margins to separate the hyperplane, sometimes at the expense of misclassifying data. The kernel function is used to transform the hyperspace and it can be a polynomial, a radial

basis function (RBF), or have a linear form. Whereas, the gamma represents the bias. Optimum algorithm parameters for the manual TC interventions and ADAS are computed respectively as: the constant C [10 (TC), 1 (ADAS)], the kernel function is set as RBF for both, and the gamma [0.1 (TC), 0.4 (ADAS)]. A smaller constant C of the ADAS model compared to the manual TC intervention model indicates higher statistical noise. Since our measurement tool is highly reliable, we attribute the noise to the complexity represented by the dataset as supported by the domain experts at INFRABEL.

In Table 3-2, we present the prediction error indices within the test set for the four ML algorithms as well as for the MLR. Recall that the data used in this paper are obtained from a highly complex STS and the relationships among the variables are highly nonlinear. We observe that compared to the ML algorithms, the multi-linear regression has the highest MAE and RMSE for both the TC and the ADAS models. This is expected and it clearly demonstrates the issues associated with approaching a complex problem with an a priori assumption of linear behavior. Notice that, the mean of efficiency score of TC and ADAS are 0.19 and 0.37, thus a difference of 0.01 in the MAE indicates a significant gap in model performance. SVM performs relatively worse than other ML algorithms and we attribute this to the sensitivity of the SVM to the selection of the kernel and hyper parameters. We experimented with the RBF kernel based on the assumption that its non-linear structure would allow for better representation of the characteristics of the STS under study. However, the RBF kernel is documented to be prone to overfitting issues, which could lead to the low performance documented in Table 3-2. The ensemble methods, RF and Xgboost, have the lowest MAE and RMSE and highest R^2 scores. This indicates low variation in both error measures demonstrating for both cases, robust and stabilized performance. The RF is clearly the best performer for both the manual TC intervention and the ADAS models with considerably less MAE and RMSE and higher R^2 scores. We attribute this to the structure of the RF algorithm that randomly generates a large number of trees to eliminate the bias of individual trees. In terms of the ANN (MLP) algorithm, we observe that its performance is only slightly worse than the ensemble methods. However, we take notice that the standard deviation in MAE is considerably higher than the other algorithms, indicating the instability of the ANN model. When looking at the R^2 scores, we clearly observe that all four ML algorithms exhibit high variability in predicting the efficiency

scores. Nevertheless, the MLR exhibits a much lower predictive power. Especially for the ADAS dataset, the R^2 scores for MLR is nearly zero in contrast to RF, which is 0.579. The considerable difference between the two R^2 scores is clear evidence that while not perfect, ML algorithms are more appropriate at explaining the non-linear relationships between dependent (efficiency scores) and independent (Z variables) variables. This ability could be leveraged with future research.

To summarize, we conclude that the ensemble methods such as the RF and the Xgboost perform remarkably well and this could be an indicator of their value for the efficiency measurement of highly complex STSs. More importantly, the results of the ML algorithms clearly outperform traditional linear regression algorithms, indicating the inadequacy of linear models to understand complexity. Lastly, not all contextual factors that could influence the workload allocation were included in this study due to framing, identification, and measurement issues[12]. We would expect the incorporation of additional contextual variables would reduce prediction errors. We next discuss the relative influence of the contextual variables on TC preferences.

Table 3-2 Prediction Accuracy of Implemented Algorithms

Error	**RF**		**Xgboost**		**SVM**		**ANN (MLP)**		**MLR**	
	TC	ADS	TC	ADS	TC	ADS	TC	ADS	TC	ADS
MAE	0.067 ±0.003	0.100 ±0.004	0.069 ±0.001	0.091 ±0.003	0.078 ±0.003	0.120 ±0.002	0.070 ±0.007	0.116 ±0.008	0.081 ±0.004	0.130 ±0.007
RMSE	0.094 ±0.006	0.137 ±0.006	0.096 ±0.005	0.125 ±0.006	0.105 ±0.006	0.166 ±0.004	0.099 ±0.006	0.152 ±0.009	0.113 ±0.006	0.183 ±0.009
R^2	0.668 ±0.023	0.579 ±0.029	0.663 ±0.040	0.569 ±0.046	0.590 ±0.038	0.240 ±0.031	0.622 ±0.052	0.371 ±0.056	0.509	0.08

As discussed in §3.3, we use the permutation importance approach for all algorithms. To recall, this approach relies on the difference in the R^2 for each individual contextual variable when compared to a permutation that iteratively reshuffles the values of the

[12] For an elaborate discussion and a demonstration of how the microeconomic production theory could be used to understand these factors please refer to Topcu, et al. (2019).

variable. The relative importance of predictor (contextual) variables in determining the manual TC and ADAS workloads is provided in Figure 3-7. Figure 3-7 indicates that dense traffic is the primary driver of the manual TC workload, followed by the traffic complexity. We observe that age, fatigue level (two "social" variables), as well as train delays have relatively minor influence on the manual workload and their respective rankings are quite similar. We attribute this to the parsimonious nature of our workload measurement model since it does not incorporate considerations related to manual TC errors and the quality of work.

In the case of the ADAS workload, we interestingly observe that the traffic complexity is the leading factor by explaining roughly 50% of the efficiency variation. It is followed by traffic density and train delays. Compared to the manual workload, we observe that delays have a significant influence. This could be explained based on on-site observations of the daily operations of the infrastructure. Since the ADAS can be activated per train, many TCs utilize the ADAS for handling routinely running sections while they manually focus on the trains and nodes with conflicts. Therefore, the delays in the system create a backlog through the infrastructure and require an increased amount of adaptations decisions. Since the adaptation decisions can only be made manually, trains that are running in no-conflict zones are handled by the ADAS. We observe that the social variables age and fatigue levels do not have a significant impact, however, both have a stronger influence on the ADAS workload compared to the manual workload. We attribute this to the state-of-the-art training procedures and the discipline of the TCs at INFRABEL. As such, these factors do not alter the daily operation of the TCs, regardless of age and mental fatigue. In other words, daily operations are influenced heavily by external factors rather than uncontrollable individual characteristic differences among TCs.

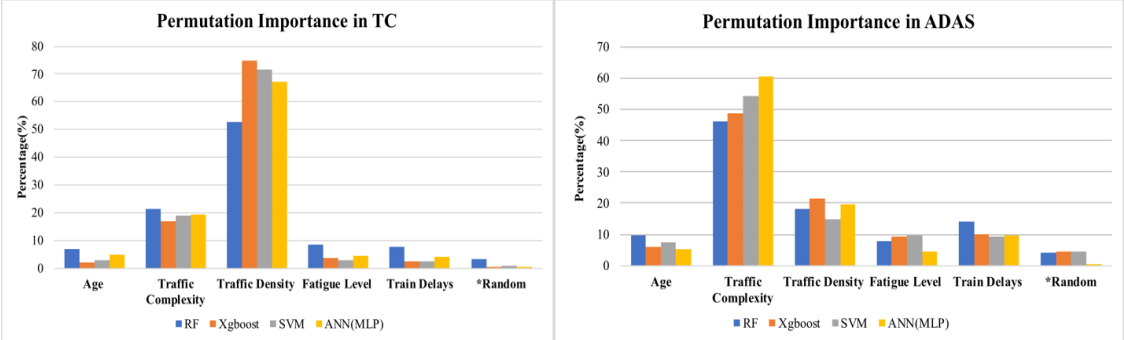

Figure 3-7 Permutation Importance of the Contextual Variables on Efficiency Scores

While the permutation importance approach captures the relative influences of contextual variables, one of its limitations is that it does not inform us about the direction of the influence. In order to demonstrate how ML techniques can be utilized to extract complex, intertwined, and non-linear relationships that cannot be properly assessed with simple regression-based methods we provide Table 3-3. Table 3-3 shows the results of the MLR regression along with the permutation importance scores for the mediocre (ANN) and the best performing ML algorithm (RF) for both the manual TC and ADAS workloads.

Table 3-3 Comparison of the Influence of Contextual Variables – MLR vs. ML

Contextual Variable	**MLR**			**ANN (MLP)**		**RF**	
		TC $\beta_0 = 0.1752 \pm 0.0006$	**ADAS** $\beta_0 = 0.4596 \pm 0.009$	**TC**	**ADS**	**TC**	**ADS**
Z1 Age	Coefficient	-0.0002	0.0005	0.001	0.003	0.118	0.229
	Std error	8.12e-05	0.000				
	P-Val[13]	0.001	0.000				
Z2 Traffic Complexity	Coefficient	6.923e-07	-2.962e-05	0.006	0.031	0.360	1.092
	Std error	5.06e-07	8.1e-07				
	P-Val	0.171	0.000				
Z3 Traffic Density	Coefficient	0.0001	-1.748e-06	0.021	0.010	0.896	0.429
	Std error	8.03e-07	1.29e-06				
	P-Val	0.000	0.174				
Z4 Fatigue Level	Coefficient	-0.08	-0.1079	0.001	0.002	0.143	0.187
	Std error	0.005	0.008				
	P-Val	0.000	0.000				
	Coefficient	-3.995e-07	-9.336e-06				

[13] Permutation importance test that was applied to the machine learning algorithms do not yield a p-value thus it is not included in the table.

Z5 Train Delays	Std error	7.02e-07	1.12e-06	0.001	0.005	0.130	0.333
	P-Val	0.569	0.000				
Random	-	-	-	0.000	0.000	0.058	0.097
SUM	-	-	-	0.031	0.051	1.704	2.367

We first interpret Table 3-3 from the 2-stage perspective as we wish to understand if our high fidelity data obtained from a fully operational STS violates fundamental assumptions of the MLR, which are: (i) the presence of a linear relationship between the independent variable and independent variables, (ii) multivariate normality, (iii) no multi-collinearity, and (iv) homoscedasticity. If all four assumptions are satisfied, then one could argue that regression based approaches could be utilized to explain the influence of the contextual variables on the efficiency score for complex systems that are similar to ours.

To test assumption 1, we check the P-values for all coefficients. A low P-value (<0.05) indicates that a predictor is an essential addition to the model. Table 3-3 indicates that, for the manual TC interventions model, the Traffic Complexity and Train Delays are insignificant whereas for the ADAS model the Traffic Density is labeled as insignificant in terms of the explaining the efficiency scores. However, results of the ML algorithms indicate exactly the opposite as these are the considerable influencers of workload allocation.

To test the second assumption, we check for the Omnibus test that captures the skewness and kurtosis of the residuals, representing whether the residuals represent a normal distribution. The Omnibus test yields considerably large values of 7,301.59 and 2,640.12 for the manual TC intervention and the ADAS models respectively. Thus, we conclude that the second assumption is violated. Next, we compute the condition number to test for the third assumption. If multi-collinearity is observed, one could expect high fluctuations to small changes in the data, commonly below 30. Both models show a high condition number of 1.94e+04, confirming that there is multi-collinearity.

Finally, we provide the scatterplot of residuals versus predicted values in Figure 3-8 for both the manual TC intervention and the ADAS models. Figure 3-8 clearly demonstrates that the data are heteroscedastic, as the scatterplot of residuals versus predicted values have a clear pattern of distribution, indicating the violation of rule four.

In conclusion, the results demonstrate that nearly every fundamental assumption of the MLR does not hold. Hence**, we argue that one should take special care when applying linear regression methods to explain the influence of Z variables on the performance of complex sociotechnical systems**. Even though we recognize that one could satisfy certain fundamental assumptions by transforming the data, this would hinder the ability to interpret the influence of contextual variables on the distribution of STS tasks represented by the efficiency scores, which is one of the primary goals of this paper.

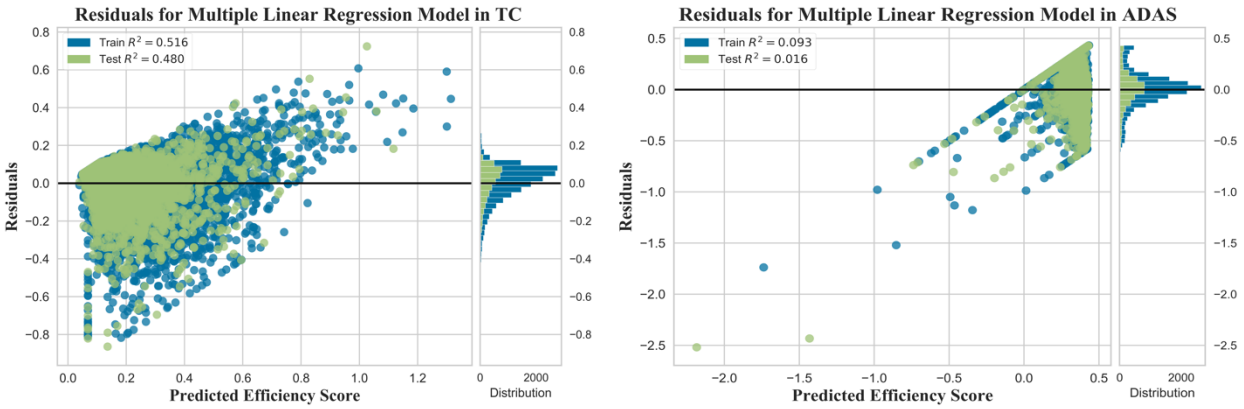

Figure 3-8 Distribution of MLR Residuals

3.5 Conclusions

Autonomous systems are becoming a larger part of our daily lives, including management of infrastructure systems that provide critical services for millions of people. As documented in this research paper, highly intertwined sociotechnical relationships that govern STS performance require an elaborate understanding to make better-informed organizational decisions. While increasing levels of automation appear as a dominant cost reduction strategy from an organizational perspective, potential STS improvement strategies need to consider the extent to which functions can be automated.

The results of this interdisciplinary operational research study and its implications for each associated discipline could be summarized with the following. From a managerial point of view, this study provides a rigorous understanding of the factors that determine the use of automation by TCs. Since the developed models are capable of considering operational complexity and are compatible with an on-site Business Intelligence tool developed at INFRABEL, it allows the managers to closely monitor the workload

distribution and the TC preferences as desired from multiple perspectives (per TC, per TCC, or for a specific time period). Thus, it provides an understanding of the operational trends of the STS with a specific focus on the task delegation between manual and autonomous decision-making without the need for self-reporting methods such as surveys. This data could potentially be used to inform future organizational change plans (such as reallocating the responsibility of segments of the infrastructure over the different workstations in order to balance the workload) or to identify context specific skill improvement areas for individual TCs.

From a performance measurement point of view, our approach, for the first time, utilizes the contextual/environmental Z variables to understand how the STS task distributions are handled. More importantly, our study hints at the weakness of 2-stage approaches to understand the influence of contextual variables for complex DMUs. As such, we observe that regression methods could classify the influence of a contextual variable on the efficiency scores (such as the traffic density for the ADAS workload in Table 3-3) much differently than the ML techniques. Given the complexity associated with both the social and technological processes of STSs and the inherent non-linearities associated with the different factors that describe their efficiency performance, we expect that ML techniques will be more useful above and beyond the computational speed issues that are currently studied in the field. This perspective allows us to further establish the interdisciplinary bridge between system science and performance measurement literatures. Previous linkages between these two research domains have been explored by considering System Dynamics techniques (Vaneman and Triantis 2007) and Complex Adaptive systems (Herrera-Restrepo and Triantis 2019).

From a sociotechnical systems point of view, our approach takes the first step towards relating the economic boundary that is driven by high level organizational decisions to the Controller workload boundary that determines daily operations. Given that future systems will rely heavily on automation supervised by specifically trained individuals, our approach concurrently helps to identify areas of improvement for the ADAS that could be used as future system design requirements. For example, the simple bar plot provided in Figure 3-7, indicates that dense traffic renders the use of automation undesirable for manual TC

interventions. Thus, ADAS improvement efforts could focus on improving performance under dense traffic. Finally, the identification of how decision-makers collaborate with automated decision-making technologies and how they distribute their workload (manual vs. automation) demonstrates under which conditions the STS operates as an engineered/automated system (that does not require human supervision to make decisions) versus when it operates as a social system (that relies heavily on human supervision and on the collaboration with the social networks). This information could be used to formulate user models to inform the design of future automated systems (Topcu and Mesmer 2018).

We realize that our research approach requires that we address multiple issues in the future. For example, continued research is required to establish the conceptual connections between ML techniques and efficiency measurement especially the physical interpretation of the various ML algorithmic parameters in the context of the STS production processes. Further, we need to consider the concurrent modeling for both the manual and automated workloads. This will require further investigation as to how these two processes are interdependent. Additionally, we believe that the understanding of the sociotechnical characteristics that influence TC preferences regarding the use of automation would allow the organization to identify future ADAS improvement areas and provide a benchmark for future organizational change decisions such as TCC mergers. In general, however, we will need to continue to investigate the decision-making and policy implications of our analysis. In general, our research provides a demonstration of regression-based models' inability to deal with complexity. On the other hand, ML approaches are intricate and transform the variables in multiple ways to establish relative influences. Thus, while it is not possible to deduct simple explanatory relationships between contextual variables and efficiency scores through permutation importance, we believe this mirrors the interdependent characteristics of the STS under study. This conclusion needs to be further established. Finally, we anticipate that the dataset contains outliers. Since this paper is concerned with understanding how much workload can be handled by the system safely, we purposefully include all observations in our analysis rather than discarding those that are different. Given that our observations are recorded from a highly reliable measurement tool and represent real instantiations of STS performance during which no accidents occurred, we consider it necessary to include all observations to be able to capture the nature of the STS.

Nevertheless, outlier analysis (Seaver and Triantis 1992) in conjunction with the approach provided in this paper will be pursued in the future. It should be noted that this additional analysis is not trivial (Herrera-Restrepo, et al. 2016).

We provided a parsimonious quantification of human and autonomous system workloads using micro-economic production theory and a unique real-world dataset. Given that increasing efficiency performance is a natural endeavor in the study of systems, we anticipate that more researchers from the systems engineering community will become exposed to its potential. We used several machine learning techniques to demonstrate how certain stakeholder preferences could be captured without the need for self-reporting methods. One could increase the prediction accuracy of ML algorithms by increasing the size of the data and including additional contextual variables. While ML is a quite powerful tool, we believe it should be considered as a complementary approach to traditional interview-based methods. Ultimately, machine prediction like other forms of prediction is limited to the variables included in the analysis. Given the highly interdependent nature of STSs, open-ended interviews could reveal a lot more information than automated algorithms. Nevertheless, this paper presents one of the first demonstrations of the combined analytic power that Data Envelopment Analysis and Machine Learning could offer, and demonstrates the need for approaching complex and non-linear phenomena with appropriate techniques.

Acknowledgements

All intellectual material discussed in this paper are protected under the non-disclosure agreement between Virginia Tech and INFRABEL. We would like to thank Dr. Renaat van de Kerkhove, Dr. Alex Fletcher, Leslie Steen, and Kristof van der Strieckt from INFRABEL for preparing the data and assisting our research with their feedback. The views expressed in this paper are those of the authors and do not necessarily reflect the opinions of INFRABEL.

References

Acemoglu, Daron, and Pascual Restrepo. 2017. “Robots and Jobs: Evidence from US Labor Markets.” NBER Working Paper No. W23285. https://ssrn.com/abstract=2941263.

Ali, Jehad, Rehanullah Khan, Nasir Ahmad, and Imran Maqsood. 2012. “Random Forests and Decision Trees.” International Journal of Computer Science Issues (IJCSI); Mahebourg 9 (5): 272–78.

Alpaydin, Ethem. 2009. Introduction to Machine Learning. MIT press.

Altmann, André, Laura Tolosi, Oliver Sander, and Thomas Lengauer. 2010. “Data and Text Mining Permutation Importance: A Corrected Feature Importance Measure.” Bioinformatics 26 (10): 1340–47. https://doi.org/10.1093/bioinformatics/btq134.

Athanassopoulos, Antreas D., and Stephen P. Curram. 1996. “A Comparison of Data Envelopment Analysis and Artificial Neural Networks as Tools for Assessing the Efficiency of Decision Making Units.” The Journal of the Operational Research Society 47 (8): 1000–1016. https://doi.org/10.2307/3010408.

Azadeh, Ali, Morteza Saberi, Reza Tavakkoli Moghaddam, and Leili Javanmardi. 2011. “An Integrated Data Envelopment Analysis–Artificial Neural Network–Rough Set Algorithm for Assessment of Personnel Efficiency.” Expert Systems with Applications 38 (3): 1364–73. https://doi.org/10.1016/j.eswa.2010.07.033.

Babajide Mustapha, Ismail, and Faisal Saeed. 2016. “Bioactive Molecule Prediction Using Extreme Gradient Boosting.” Molecules 21 (8): 983. https://doi.org/10.3390/molecules21080983.

Banker, Rajiv D., Abraham Charnes, and William Wager Cooper. 1984. “Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis.” Management Science 30 (9): 1078–1092.

Banker, Rajiv D., and Richard C. Morey. 1986. “The Use of Categorical Variables in Data Envelopment Analysis.” Management Science 32 (12): 1613–1627.

Banker, Rajiv D., and Ram Natarajan. 2008. “Evaluating Contextual Variables Affecting Productivity Using Data Envelopment Analysis.” Operations Research 56 (1): 48–58. http://www.jstor.org.ezproxy.lib.vt.edu/stable/25147166.

Banker, Rajiv, Ram Natarajan, and Daqun Zhang. 2019. “Two-Stage Estimation of the Impact of Contextual Variables in Stochastic Frontier Production Function Models Using Data Envelopment Analysis: Second Stage OLS versus Bootstrap Approaches.” European Journal of Operational Research, Advances in Data Envelopment Analysis, 278 (2): 368–84. https://doi.org/10.1016/j.ejor.2018.10.050.

Beehr, Terry A. 2014. Psychological Stress in the Workplace (Psychology Revivals). Routledge.

Bergstra, James, and Yoshua Bengio. 2012. “Random Search for Hyper-Parameter Optimization.” Journal of Machine Learning Research 13 (Feb): 281–305.

Borchani, Hanen, Gherardo Varando, Concha Bielza, and Pedro Larrañaga. 2015. “A Survey on Multi-Output Regression.” Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 5 (5): 216–233.

Boser, Bernhard E, Isabelle M Guyon, and Vladimir N Vapnik. 1992. “A Training Algorithm for Optimal Margin Classiiers.” Proceedings of the Fifth Annual Workshop on Computational Learning Theory ACM.

Breiman, Leo. 2001a. “Random Forests.” Machine Learning 45 (1): 5–32.

———. 2001b. “Random Forests.” Machine Learning 45 (1).

———. 2017. Classification and Regression Trees. Routledge. https://doi.org/10.1201/9781315139470.

Chambers, Christopher P., and Federico Echenique. 2016. Revealed Preference Theory. Vol. 56. Cambridge University Press.

Charnes, Abraham, William W. Cooper, and Edwardo Rhodes. 1978. “Measuring the Efficiency of Decision Making Units.” European Journal of Operational Research 2 (6): 429–444. http://www.sciencedirect.com/science/article/pii/0377221778901388.

Chen, Tianqi, and Carlos Guestrin. 2016a. “XGBoost: A Scalable Tree Boosting System.” In Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. KDD ’16. New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785.

———. 2016b. “XGBoost: A Scalable Tree Boosting System.” Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’16, 785–94. https://doi.org/10.1145/2939672.2939785.

Chollet, Francois. 2015. Keras. https://keras.io.

Cortes, Corinna, and Vladimir Vapnik. 1995. “Support-Vector Networks.” Machine Learning 20 (3): 273–97.

Costa, Álvaro, and Raphael N. Markellos. 1997. “Evaluating Public Transport Efficiency with Neural Network Models.” Transportation Research Part C: Emerging Technologies 5 (5): 301–12. https://doi.org/10.1016/S0968-090X(97)00017-X.

Daraio, Cinzia, and Léopold Simar. 2005. “Introducing Environmental Variables in Nonparametric Frontier Models: A Probabilistic Approach.” Journal of Productivity Analysis 24 (1): 93–121. https://doi.org/10.1007/s11123-005-3042-8.

Dawson, Drew, and Kirsty McCulloch. 2005. “Managing Fatigue: It’s about Sleep.” Sleep Medicine Reviews 9 (5): 365–380.

Debreu, Gerard. 1951. “The Coefficient of Resource Utilization.” Econometrica 19 (3): 273–92. https://doi.org/10.2307/1906814.

Farrell, M. J. 1957. “The Measurement of Productive Efficiency.” Journal of the Royal Statistical Society. Series A (General) 120 (3): 253–90. https://doi.org/10.2307/2343100.

Ferguson, Sally A., Nicole Lamond, Katie Kandelaars, Sarah M. Jay, and Drew Dawson. 2008. “The Impact of Short, Irregular Sleep Opportunities at Sea on the Alertness of Marine Pilots Working Extended Hours.” Chronobiology International 25 (2–3): 399–411.

Folkard, Simon, Karen A. Robertson, and Mick B. Spencer. 2007. “A Fatigue/Risk Index to Assess Work Schedules.” Somnologie-Schlafforschung Und Schlafmedizin 11 (3): 177–185.

Franssen, Maarten. 2005. “Arrow’s Theorem, Multi-Criteria Decision Problems and Multi-Attribute Preferences in Engineering Design.” Research in Engineering Design 16 (1–2): 42–56. https://doi.org/10.1007/s00163-004-0057-5.

Frey, Carl Benedikt, and Michael A. Osborne. 2017. “The Future of Employment: How Susceptible Are Jobs to Computerisation?” Technological Forecasting and Social Change 114: 254–280.

Friedman, Jerome H. 2001. “Greedy Function Approximation: A Gradient Boosting Machine.” The Annals of Statistics 29 (5): 1189–1232. https://doi.org/10.1214/aos/1013203451.

———. 2002. “Stochastic Gradient Boosting.” Computational Statistics & Data Analysis 38 (4): 367–78. https://doi.org/10.1016/S0167-9473(01)00065-2.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2000. “Additive Logistic Regression: A Statistical View of Boosting (With Discussion and a Rejoinder by the Authors).” The Annals of Statistics 28 (2): 337–407. https://doi.org/10.1214/aos/1016218223.

Gardner, M W, and S R Dorling. 1998. “ARTIFICIAL NEURAL NETWORKS (THE MULTILAYER PERCEPTRON)-A REVIEW OF APPLICATIONS IN THE ATMOSPHERIC SCIENCES.” Atmospheric Environment. Vol. 32.

Grundgeiger, Tobias, Penelope M. Sanderson, and R. Key Dismukes. 2015. “Prospective Memory in Complex Sociotechnical Systems.” Zeitschrift Für Psychologie.

Guyon, Isabelle, B Boser, and Vladimir Vapnik. 1993. “Automatic Capacity Tuning of Very Large VC-Dimension Classifiers.” In Advances in Neural Information Processing Systems, 147–55.

Hartigan, J. A., and M. A. Wong. 1979. “Algorithm AS 136: A K-Means Clustering Algorithm.” Journal of the Royal Statistical Society. Series C (Applied Statistics) 28 (1): 100–108. https://doi.org/10.2307/2346830.

Herrera-Restrepo, Oscar, and Konstantinos Triantis. 2019. “Enterprise Design through Complex Adaptive Systems and Efficiency Measurement.” European Journal of Operational Research, Advances in Data Envelopment Analysis, 278 (2): 481–97. https://doi.org/10.1016/j.ejor.2018.12.002.

Herrera-Restrepo, Oscar, Konstantinos Triantis, William L. Seaver, Joseph C. Paradi, and Haiyan Zhu. 2016. “Bank Branch Operational Performance: A Robust Multivariate and Clustering Approach.” Expert Systems with Applications 50: 107–119. http://www.sciencedirect.com/science/article/pii/S0957417415008271.

Hoff, Ayoe. 2007. “Second Stage DEA: Comparison of Approaches for Modelling the DEA Score.” European Journal of Operational Research 181 (1): 425–35. https://doi.org/10.1016/j.ejor.2006.05.019.

Hornik, Kurt, Maxwell Stinchcombe, and Halbert White. 1989. “Multilayer Feedforward Networks Are Universal Approximators.” Neural Networks 2 (5): 359–66. https://doi.org/10.1016/0893-6080(89)90020-8.

Hubert, Mia, Peter J. Rousseeuw, and Karlien Vanden Branden. 2005. “ROBPCA: A New Approach to Robust Principal Component Analysis.” Technometrics 47 (1): 64–79. https://doi.org/10.1198/004017004000000563.

Jain, Anil K., and Jianchang Mao. 1996. “Artificial Neural Networks: A Tutorial.” Computer 29 (3).

Jain, Anil K., Jianchang Mao, and K. M. Mohiuddin. 1996. “Artificial Neural Networks: A Tutorial.” Computer, no. 3: 31–44.

Johnson, Andrew L., and Timo Kuosmanen. 2011. “One-Stage Estimation of the Effects of Operational Conditions and Practices on Productive Performance: Asymptotically Normal and Efficient, Root-<Emphasis Type="Italic">n</Emphasis> Consistent StoNEZD Method.” Journal of Productivity Analysis 36 (2): 219–30. https://doi.org/10.1007/s11123-011-0231-5.

———. 2012. “One-Stage and Two-Stage DEA Estimation of the Effects of Contextual Variables.” European Journal of Operational Research 220 (2): 559–70. https://doi.org/10.1016/j.ejor.2012.01.023.

Karlaftis, M. G., and E. I. Vlahogianni. 2011. “Statistical Methods versus Neural Networks in Transportation Research: Differences, Similarities and Some Insights.” Transportation Research Part C: Emerging Technologies 19 (3): 387–99. https://doi.org/10.1016/j.trc.2010.10.004.

Keeney, Ralph L., and Howard Raiffa. 1993. Decisions with Multiple Objectives: Preferences and Value Trade-Offs. Cambridge University Press.

Khezrimotlagh, Dariush, Joe Zhu, Wade D. Cook, and Mehdi Toloo. 2019. “Data Envelopment Analysis and Big Data.” European Journal of Operational Research 274 (3): 1047–54. https://doi.org/10.1016/j.ejor.2018.10.044.

Kleiner, Brian M., Lawrence J. Hettinger, David M. DeJoy, Yuang-Hsiang Huang, and Peter E. D. Love. 2015. “Sociotechnical Attributes of Safe and Unsafe Work Systems.” Ergonomics 58 (4): 635–49. https://doi.org/10.1080/00140139.2015.1009175.

Kohavi, Ron. 1995. “A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection.” In Ijcai, 14:1137–1145. Montreal, Canada.

Koopmans, Tjalling C. 1951. An Analysis of Production as an Efficient Combination of Activities. Cowles Commission for Research in Economics. New York: John Wiley & Sons.

Kotsiantis, Sotiris B., I. Zaharakis, and P. Pintelas. 2007. “Supervised Machine Learning: A Review of Classification Techniques.” Emerging Artificial Intelligence Applications in Computer Engineering 160: 3–24.

Kroes, Peter, Maarten Franssen, Ibo van de Poel, and Maarten Ottens. 2006. “Treating Sociotechnical Systems as Engineering Systems: Some Conceptual Problems.” Systems Research and Behavioral Science 23 (6): 803–814. http://onlinelibrary.wiley.com/doi/10.1002/sres.703/full.

Leveson, Nancy G. 2011. “Applying Systems Thinking to Analyze and Learn from Events.” Safety Science 49 (1): 55–64. http://www.sciencedirect.com/science/article/pii/S0925753510000068.

Lippmann, Richard P. 1987. “An introduction to Computing with Neural Nets,”.” IEEE Assp Magazine 4 (2): 4–22.

McCulloch, Warren S., and Walter Pitts. 1943. “A Logical Calculus of the Ideas Immanent in Nervous Activity.” The Bulletin of Mathematical Biophysics 5 (4): 115–33. https://doi.org/10.1007/BF02478259.

McDonald, John. 2009. “Using Least Squares and Tobit in Second Stage DEA Efficiency Analyses.” European Journal of Operational Research 197 (2): 792–98. https://doi.org/10.1016/j.ejor.2008.07.039.

Mountrakis, Giorgos, Jungho Im, and Caesar Ogole. 2011. “Support Vector Machines in Remote Sensing: A Review.” ISPRS Journal of Photogrammetry and Remote Sensing 66 (3): 247–59. https://doi.org/10.1016/j.isprsjprs.2010.11.001.

National Transportation Safety Board. 2016. “Amtrak Train Collision with Maintenance-of-Way Equipment, Chester, Pennsylvania, April 3, 2016.” Accident Report NTSB/RAR-17/02. Washington DC: NTSB. https://www.ntsb.gov/investigations/AccidentReports/Reports/RAR1702.pdf.

Nørgård, Peter Magnus, Ole Ravn, Niels Kjølstad Poulsen, and Lars Kai Hansen. 2000. “Neural Networks for Modelling and Control of Dynamic Systems-A Practitioner’s Handbook.”

O’Donnell, Christopher J., D. S. Prasada Rao, and George E. Battese. 2008. “Metafrontier Frameworks for the Study of Firm-Level Efficiencies and Technology Ratios.” Empirical Economics 34 (2): 231–55. https://doi.org/10.1007/s00181-007-0119-4.

Oshiro, Thais Mayumi, Pedro Santoro Perez, and José Augusto Baranauskas. 2012. “How Many Trees in a Random Forest?” In Machine Learning and Data Mining in Pattern Recognition, edited by Petra Perner, 154–68. Lecture Notes in Computer Science. Springer Berlin Heidelberg.

Oster, Sharon M. 1999. Modern Competitive Analysis. 3rd ed. Oxford University Press.

O’Sullivan, Arthur, and Steven M Sheffrin. 2007. Economics: Principles in Action. Boston, MA: Pearson/Prentice Hall.

Pachl, Joern. 2002. Railway Operation and Control.

Paltrinieri, Nicola, Louise Comfort, and Genserik Reniers. 2019. “Learning about Risk: Machine Learning for Risk Assessment.” Safety Science 118 (October): 475–86. https://doi.org/10.1016/j.ssci.2019.06.001.

Paradi, Joseph C., and H. David Sherman. 2014. “Seeking Greater Practitioner and Managerial Use of DEA for Benchmarking.” Data Envelopment Analysis Journal 1 (1): 29–55. https://www.researchgate.net/profile/Joseph_Paradi/publication/281010454_Seeking_Greater_Practitioner_and_Managerial_Use_of_DEA_for_Benchmarking/links/567eb75d08ae051f9ae655de.pdf.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. “Scikit-Learn: Machine Learning in Python.” J. Mach. Learn. Res. 12 (November): 2825–2830. http://dl.acm.org/citation.cfm?id=1953048.2078195.

Rasmussen, Jens. 1997. “Risk Management in a Dynamic Society: A Modelling Problem.” Safety Science 27 (2): 183–213. https://doi.org/10.1016/S0925-7535(97)00052-0.

Ray, Subhash C. 1988. “Data Envelopment Analysis, Nondiscretionary Inputs and Efficiency: An Alternative Interpretation.” Socio-Economic Planning Sciences 22 (4): 167–76. https://doi.org/10.1016/0038-0121(88)90003-1.

Roets, Bart, and Johan Christiaens. 2017. “Shift Work, Fatigue and Human Error: An Empirical Analysis of Railway Traffic Control.” Journal of Transportation Safety & Security 0 (ja): 1–18. https://doi.org/10.1080/19439962.2017.1376022.

Roets, Bart, Marijn Verschelde, and Johan Christiaens. 2018. “Multi-Output Efficiency and Operational Safety: An Analysis of Railway Traffic Control Centre Performance.” European Journal of Operational Research. https://doi.org/10.1016/j.ejor.2018.04.045.

Rosenblatt, F. 1958. “THE PERCEPTRON: A PROBABILISTIC MODEL FOR INFORMATION STORAGE AND ORGANIZATION IN THE BRAIN 1.” Psychological Review. Vol. 65.

Ruggiero, John. 1996. “On the Measurement of Technical Efficiency in the Public Sector.” European Journal of Operational Research 90 (3): 553–65. https://doi.org/10.1016/0377-2217(94)00346-7.

Salmon, Paul M., Neville A. Stanton, Guy H. Walker, Daniel Jenkins, Darshna Ladva, Laura Rafferty, and Mark Young. 2009. “Measuring Situation Awareness in Complex Systems: Comparison of Measures Study.” International Journal of Industrial Ergonomics 39 (3): 490–500. https://doi.org/10.1016/j.ergon.2008.10.010.

Salmon, Paul M., Guy H. Walker, and Neville A. Stanton. 2016. “Pilot Error versus Sociotechnical Systems Failure: A Distributed Situation Awareness Analysis of Air France 447.” Theoretical Issues in Ergonomics Science 17 (1): 64–79. https://doi.org/10.1080/1463922X.2015.1106618.

Samoilenko, Sergey, and Kweku-Muata Osei-Bryson. 2013. “Using Data Envelopment Analysis (DEA) for Monitoring Efficiency-Based Performance of Productivity-Driven Organizations: Design and Implementation of a Decision Support System.” Omega, Data Envelopment Analysis: The Research Frontier - This Special Issue is dedicated to the memory of William W. Cooper 1914-2012, 41 (1): 131–42. https://doi.org/10.1016/j.omega.2011.02.010.

Samuelson, Paul A. 1948. “Consumption Theory in Terms of Revealed Preference.” Economica 15 (60): 243–53. https://doi.org/10.2307/2549561.

Santin, Daniel, Francisco J. Delgado, and Aurelia Valino. 2004. “The Measurement of Technical Efficiency: A Neural Network Approach.” Applied Economics 36 (6): 627–635.

Scott, W. Richard. 2015. Organizations and Organizing: Rational, Natural and Open Systems Perspectives. Routledge.

Seaver, Bill L., and Konstantinos P. Triantis. 1992. “A Fuzzy Clustering Approach Used in Evaluating Technical Efficiency Measures in Manufacturing.” Journal of Productivity Analysis 3 (4): 337–363. http://www.springerlink.com/index/x05527653768225m.pdf.

Sheridan, Robert P., Wei Min Wang, Andy Liaw, Junshui Ma, and Eric M. Gifford. 2016. “Extreme Gradient Boosting as a Method for Quantitative Structure–Activity Relationships.” Journal of Chemical Information and Modeling 56 (12): 2353–60. https://doi.org/10.1021/acs.jcim.6b00591.

Simar, Léopold, and Paul W. Wilson. 2007. “Estimation and Inference in Two-Stage, Semi-Parametric Models of Production Processes.” Journal of Econometrics 136 (1): 31–64. https://doi.org/10.1016/j.jeconom.2005.07.009.

———. 2011. “Two-Stage DEA: Caveat Emptor.” Journal of Productivity Analysis 36 (2): 205. https://doi.org/10.1007/s11123-011-0230-6.

Sohn, So Young, and Tae Hee Moon. 2004. “Decision Tree Based on Data Envelopment Analysis for Effective Technology Commercialization.” Expert Systems with Applications 26 (2): 279–84. https://doi.org/10.1016/j.eswa.2003.09.011.

Song, Ma-Lin, Ron Fisher, Jian-Lin Wang, and Lian-Biao Cui. 2018. “Environmental Performance Evaluation with Big Data: Theories and Methods.” Annals of Operations Research 270 (1): 459–72. https://doi.org/10.1007/s10479-016-2158-8.

Strobl, Carolin, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. 2007. “Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution.” BMC Bioinformatics 8 (1): 25. https://doi.org/10.1186/1471-2105-8-25.

Tan, Pang-Ning. 2018. Introduction to Data Mining. Pearson Education India.

Topcu, Taylan G., and Bryan L. Mesmer. 2018. “Incorporating End-User Models and Associated Uncertainties to Investigate Multiple Stakeholder Preferences in System Design.” Research in Engineering Design 29 (3): 411–31. https://doi.org/10.1007/s00163-017-0276-1.

Topcu, Taylan G., Konstantinos Triantis, and Bart Roets. 2019. “Estimation of the Workload Boundary in Socio-Technical Infrastructure Management Systems: The Case of Belgian Railroads.” European Journal of Operational Research 278 (1): 314–29. https://doi.org/10.1016/j.ejor.2019.04.009.

Triantis, K. 2015. “Engineering Design and Efficiency Measurement: Issues and Future Research Opportunities.” Data Envelopment Analysis Journal 1 (2): 81–112. http://econpapers.repec.org/RePEc:now:jnldea:103.00000008.

Triantis, Konstantinos, Devang Sarayia, and Bill Seaver. 2010. “Using Multivariate Methods to Incorporate Environmental Variables for Local and Global Efficiency Performance Analysis.” INFOR: Information Systems and Operational Research 48 (1): 39–52. https://doi.org/10.3138/infor.48.1.039.

Tu, Jack V. 1996. “Advantages and Disadvantages of Using Artificial Neural Networks versus Logistic Regression for Predicting Medical Outcomes.” Journal of Clinical Epidemiology 49 (11): 1225–31. https://doi.org/10.1016/S0895-4356(96)00002-9.

Vaneman, Warren K., and Konstantinos Triantis. 2007. “Evaluating the Productive Efficiency of Dynamical Systems.” Engineering Management, IEEE Transactions On 54 (3): 600–612. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4278019.

Vapnik, Vladimir, Steven E Golowich, and Alex J Smola. 1997. “Support Vector Method for Function Approximation, Regression Estimation and Signal Processing.” In Advances in Neural Information Processing Systems, 281–87.

Vapnik, Vladimir, and A J Lerner. 1963. “Generalized Portrait Method for Pattern Recognition.” Automation and Remote Control 24 (6): 774–80.

Wilson, John R. 2000. “Fundamentals of Ergonomics in Theory and Practice.” Applied Ergonomics 31 (6): 557–67. https://doi.org/10.1016/S0003-6870(00)00034-X.

Zhu, Nan, Chuanjin Zhu, Ali Emrouznejad, and Luman Chen. 2018. “A Combined DEA with Machine Learning Algorithms for Measuring the Efficiency of Chineese Listed Manufacturing Plants.” In DEA40: Data Envelopment Analysis and Performance Measurement: Recent Developments, 12–21. Birmingham, UK: Aston Business School.

Zhu, Qingyuan, Jie Wu, and Malin Song. 2018. “Efficiency Evaluation Based on Data Envelopment Analysis in the Big Data Context.” Computers & Operations Research 98: 291–300.

Chapter 4. Complementary Assessment of Efficiency Measurement under Contextual Heterogeneity: Insights from Sociotechnical Systems

Taylan G. Topcu, Konstantinos Triantis, Bart Roets, William Seaver, Cinzia Daraio, and Oscar Herrera-Restrepo

Abstract

The efficiency measurement literature investigates the performance of comparable decision-making units (DMUs). In this literature, the term homogeneity is used to represent similarities in both the employed production technology and the uncontrollable contextual factors that influence the DMU's operational characteristics. While there are numerous analytical techniques that are concerned with heterogeneous production possibility sets, these fundamentally differ in terms of their representations of the transformation process, assumptions, limitations, mathematical structure, and consequently, their practical usefulness. This paper provides an outline of methods that measure efficiency under contextual heterogeneity and presents an application of two alternative approaches by focusing on their complementary insights. These methods include a robust multi-variate and an 2-stage approach. Depending on whether the separability assumption holds, the 2-stage approach could either follow a Simar and Wilson (2007) strategy or could compute a conditional efficiency measure (Badin, Simar, and Daraio 2012). Considering that many application studies can be constrained by data availability, we explore the sensitivity of the investigated approaches to sample size. We ensure the validity of the obtained insights by conducting our study using on data from Belgian National railway organization's fully operational traffic control centers, and discuss the results with domain experts. We conclude with a taxonomy of complementary and contrasting roles. Our study indicates that each approach could provide a unique perspective in terms of explaining the transformation processes. However, the insights need to be interpreted with caution, especially considering the fundamental assumptions that enable each method.

Keywords: Data Envelopment Analysis (DEA), multivariate method, two-stage method, contextual variables, environmental factors, conditional efficiency, sociotechnical systems.

4.1 Introduction

Data Envelopment Analysis (DEA) (Farrell 1957; Charnes, Cooper, and Rhodes 1978) investigates the productive efficiency of decision-making units (DMUs) from a relative perspective meaning that the performance of a unit is computed with respect to the performance of its peers. Since the approach is of relative nature, the framing of the production possibility set (PPS, also represented in this paper as Ψ) has a predominant influence on the outcome of individual efficiency scores. A fundamental assumption of the Farrell measure (Farrell 1957) of the Pareto-Koopmans frontier (Koopmans 1951; Charnes et al. 1985) is the *comparability* or *the homogeneity assumption*. The *homogeneity assumption* argues that the investigated DMUs that constitute the PPS have to be comparable in terms of purpose, employed production technologies, and the uncontrollable contextual and environmental factors that delineate their performance environments. The issue with the homogeneity assumption is that, unlike the axioms of production (Färe and Grosskopf 2012, 11–16), it is not explicitly defined as a formal mathematical construct in the literature thus adherence to this assumption is tested by heuristics. Consequently, many efficiency studies either disregard this fundamental assumption or only consider it using a qualitative rule of thumb perspective.

The primary concern of this paper is that, when considered in detail, many real-world transformation processes experience high variations in terms of the contextual factors that influence their operation. Meaning that the scope, magnitude, and the distribution of contextual factors could vary considerably among the DMUs. We define such PPSs as heterogeneous and argue that if such instances are not treated appropriately (by taking the contextual (Z) variables into account), the insights of the DEA approach might lose its validity. There are numerous analytical methods that are concerned with performance measurement in heterogeneous PPSs and there is an ongoing debate in the literature regarding their utility, limitations, mathematical rigor, and validity (Simar and Wilson 2011; Dai and Kuosmanen 2014; Daraio, Simar, and Wilson 2018). Additionally, practical usefulness of these approaches are subject to sample size and data availability issues (Dyson et al. 2001). Given these considerations, this paper provides an outline based on the literature of the strengths and limitations of efficiency measurement methods that are

concerned with heterogeneous PPSs. We then identify two leading methods based on their applicability to a wide-variety of efficiency measurement scenarios and discuss the implementation of these methods. More specifically we consider the multivariate clustering approach (Triantis, Seaver, and Sarayia 2010) and a two stage method that could either follow the separability assumption (Simar and Wilson 2007) or provide a conditional efficiency measure (Cazals, Florens, and Simar 2002; Daraio and Simar 2005). We provide an in-depth discussion of why we consider these methods appropriate in §2 and document how their interpretation of the efficiency performance of a complex sociotechnical infrastructure management system will vary with respect to the size of available data.

To ensure the insights of this paper are verified and to demonstrate how certain transformation processes could be subject to high levels of contextual variation, we conduct this research with INFRABEL, the Belgian National Railroad company. INFRABEL manages its infrastructure from Traffic Control Centers (TCCs). TCCs are sociotechnical systems (STSs) that operate through a collaboration of human decision-makers known as Controllers with autonomous systems. Since the TCCs make safety-critical operational decisions[14] to sustain railroad transportation services 24/7, their operational context is subject to highly dynamic uncontrollable factors (e.g., traffic, mental fatigue, delays in schedule, etc.). In other words, STSs such as TCCs demonstrate heterogeneity of their PPSs and consequently the variation in the contextual (Z) variables needs to be accounted for. This academia-industry collaboration allows us to have access to a unique dataset that provides disaggregate measurements of Controller actions on an hourly level. The data that enables this study differs from previous research (Topcu, Triantis, and Roets 2019) in terms of its: (i) temporal and spatial range, as it spans one month of observations from nine different TCCs, (ii) sample size, as the data includes over 40k observations, and its (iii) addition of new variables including the train delays, phone calls, and anticipation tool usage (described in detail in §3.1).

[14] In the safety science literature, safety critical decisions are those that if taken incorrectly will lead to severe and many times, disastrous outcomes.

There are two main contributions of this paper. The first is documenting the differences of the two approaches when considering the Z variables. We achieve this through a comprehensive exploration of their complementary insights within a fully operational STS. We consider this an important research task because there is a lack of consensus regarding the usefulness and validity of the results obtained from these methods. This paper aims to fill this gap through a comprehensive empirical study that obtains verification from domain experts. Moreover, since the application area is a fully operational infrastructure management system, this paper also serves to spread the use of efficiency measurement for complex systems, which from an application perspective remains a long-desired goal for our research community (Paradi & Sherman, 2014; Triantis, 2015). Considering that increasingly complex and automated management systems will dominate the future of our society (Acemoglu and Restrepo 2017; Frey and Osborne 2017), the management of autonomous infrastructures is an important research area to explore for the efficiency measurement community.

The second contribution of this paper is to demonstrate the sensitivity of results and insights as they pertain to the sample size. We investigate this issue because application studies are often constrained by the size and dimension of the available data. Therefore, we present our comparison on two datasets. The first is the universe that spans across nine different TCCs. The second dataset is a subset of the universe that only includes observations from a single TCC. Both datasets are for the duration of a random weekday, and are composed of observations that represent one hour of traffic control activities. To ensure high fidelity, we adopt the following validation strategy. First, we obtain the data from a single high precision measurement source thus effectively eliminating potential measurement inconsistencies and errors (Krantz et al. 1971). Second, we establish construct validity, as it is represented by the validity of our measurement instrument (Broniatowski and Tucker 2017), by basing our DEA model formulation on a recently published article (Topcu, Triantis, and Roets 2019). Third, we verify the results of both approaches on-site, on a fully-operational STS, using large datasets, without the need for simulated test conditions, data generating processes, or other restrictive statistical assumptions. Finally, we seek validation directly from INFRABEL domain experts based on their knowledge of the process and interpretation of the operational events. We then

present a structured taxonomy that includes the assumptions, treatment of contextual variables, computation time, robustness, sensitivity to sample size, limitations, and managerial information of both approaches. By doing so, our research provides a benchmark for current modeling approaches based on realistic evidence and focusing on their practical usefulness and their ability to interpret the operational realities of complex production process.

This study is structured in the following manner. Section 4.2 provides a brief literature review. Section 4.3 describes the data, compares the methodologies, and discusses model formulations. Section 4.4 presents the results including the validation with domain experts, and Section 4.5 concludes.

4.2 Literature Review

DEA's microeconomic roots and its empirical perspective (Koopmans 1951; Farrell 1957) allows to investigate a wide array of processes, however, its application areas have historically focused on agriculture, banking, supply chain management, transportation, and public policy (Paradi and Sherman 2014; Emrouznejad and Yang 2018). In order to spread the use of DEA to complex production processes (e.g., management of autonomous systems), one needs to consider the dynamic changes in its environment (Triantis 2015). Although not explicitly defined in the efficiency measurement literature, the homogeneity assumption has a strong relationship with the nature of the artificial artifacts it describes (e.g., DMUs), and the homogeneity assumption constitutes a foundational object of study in the systems science literature. According to Simon (1996), man-made entities, e.g., DMUs and PPS in DEA, can be defined by three determinants: (i) the purpose, (ii) the internal characteristics, and (iii) the environment in which the entity or artifact performs. In DEA, the purpose of a DMU corresponds to the objective of the production technology and is represented with the inputs and outputs required/generated to fulfill the mission. The internal characteristic differences and environmental influences are considered alike, as contextual factors. However, historically, many DEA studies overlooked the fact that homogeneity with respect to purpose or mission does not necessarily translate into homogeneity in relation to contextual influences.

In the following sub-sections, we will briefly cover methods proposed in the DEA literature to deal with contextual heterogeneity. We discuss each method based on their fundamental approach to heterogeneity, other similar methods in the literature, the structure of the algorithms, application areas, and their limitations.

4.2.1 The Robust Multivariate Method

The first approach we investigate is the robust multivariate method (TSS) (Triantis, Seaver and Sarayia 2010). The fundamental motivation of TSS is to ensure that the homogeneity assumption holds. To achieve this, TSS accounts for the heterogeneity of the PPS through robust multivariate tests (Hubert, Rousseeuw, and Branden 2005) and inspects if the differences among the contextual variables are drastic enough to violate the homogeneity assumption. The fundamental understanding of the TSS is that if the PPS is heterogeneous then the insights from the DEA approach are nullified. For such PPSs, TSS proposes to subset the observations into clusters in which the homogeneity assumption holds. In other words, TSS simply relies on preserving the seminal idea of Farrell that compared homogenous production units (Farrell 1957). Interestingly, a similar approach to TSS was advocated by Farrell, as he referred to the factors that determine homogeneity, the contextual Z variables, as *quasi factors* and suggested that re-arrangement of the PPS is the most obvious solution (Farrell 1957, page 259):

"The simplest and the most obvious solution to this problem (referring to differences in quasi-factors) is analogous to that of economies of scale, that is, to divide the observations into homogenous in the quasi-factor."

The earlier research supported this idea. It was argued that, to be able to properly assess the notion of efficiency, one has to quantify or at least consider the environmental differences within the PPS because the range of managerial decisions are determined by the production environment (Hall and Winsten 1959). Banker and Morey proposed arranging the peer selection criteria by considering the differences among DMUs (Banker and Morey 1986). Similarly, others proposed to exclude the DMUs that operate in more favorable environments from peer selection (Ruggiero 1996). From this perspective, the TSS provides two significant improvements: (i) it extends the motivation of these approaches to a multivariate case and (ii) it employs robust outlier identification techniques to rigorously identify the differences in the production environments. TSS relies on robust

principal component analysis (ROBPCA) (Hubert, Rousseeuw, and Branden 2005), to identify statistically influential observations within the Z variables. Practically, this step exaggerates the differences among the observations by reducing the dimension of the data to principal components that describe the bulk of the information. The robust principal components of the data are then parsed by following a nearest neighbors algorithm that captures the outlying observations by grouping them with similar observations (Wong and Lane 1983). The nearest neighbor clusters are then grouped into larger homogenous subgroups through k-means algorithm (Hartigan and Wong 1979), and the best combination of clusters is identified by minimizing the jackknife error (Miller 1974). The use of ROBPCA for the identification of influential observations allows one to classify statistically different contextual influences. This creates managerial value, by identifying the contextual operational differences. This provides considerable flexibility to the TSS approach, since by using the relatively homogenous clusters, any DEA algorithm could be utilized to evaluate the PPS from a meta vs. in-cluster perspective (O'Donnell, Rao, and Battese 2008). Other studies pursue similar goals using classification algorithms without focusing on the robustness considerations that are rigorously addressed in TSS (Dai and Kuosmanen 2014).

The application areas of the TSS approach include non-profit service organizations (Seaver and Triantis 1992; Athanassopoulos and Triantis 1998) and banking (Herrera-Restrepo, et al. 2016). A recent implementation study on infrastructure management systems (Topcu, Triantis, and Roets 2019) demonstrated an in-depth walkthrough of a TSS model formulation by considering a wide array contextual factors. These factors were documented domain experts provided face-validation. Thus, a field-proven practical strength of the TSS is its ability to quantify the aggregate impact of the contextual factors, that is represented by the technology gap (O'Donnell, Rao, and Battese 2008) or the difference between in-cluster and meta efficiency scores. Topcu, Triantis, and Roets (2019) used this approach to quantify the impact of sociotechnical work environment factors on the overall system risk levels and emphasize that disregarding contextual factors could lead up to 50% underestimation of the efficiency scores (Topcu, Triantis, and Roets 2019). Nevertheless, a leading limitation of the method is its inability to compute the individual

contribution of each contextual factor. We next describe the methods that are concerned with explaining the influence of individual contextual variables.

4.2.2 Alternative 2-Stage Approaches, the Separability Assumption and Conditional Measures

The second approach we investigate are the two-stage approaches (TSAs). TSAs are named after the structure of their algorithm. In the first stage of TSA efficiency scores are computed, and in the second stage, some form of a prediction or regression technique is used to explain the influence of contextual variables on the Pareto-Koopmans frontier. TSA assumes that, unbiased and efficient estimation of the contextual variables require a joint estimation of the frontier with the influence of contextual variables, thus the first stage efficiency score calculation needs to take this into consideration (Wang and Schmidt 2002). Numerous publications have experimented with this idea (Ray 1988; Lovell, Walters, and Wood 1994; Stanton 2002; Turner, Windle, and Dresner 2004; Chilingerian and Sherman 2004). A considerable amount of second stage regression based approaches have experimented by using tobit and ordinary least squares regression (Hoff 2007; Banker and Natarajan 2008; McDonald 2009). These approaches were criticized to be inconsistent in terms of the statistical meaning of their second stage regression results (Simar and Wilson 2007; 2011). In their seminal 2007 publication, Simar and Wilson (SW) argued that for a second stage regression to be statistically consistent and meaningful, the PPS should satisfy a restrictive hypothesis known as the separability condition. SW is a popular TSA method, with over 2,275 citations on Google Scholar (Barros, Nektarios, and Assaf 2010; Blank and Valdmanis 2010; Latruffe, Davidova, and Balcombe 2008). Therefore, we explore SW in our study and proceed to discuss the separability condition.

In simple terms, the separability condition or assumption, argues that the position and the shape of the frontier, that represents the production technology, is independent from the contextual factors. This suggests that the contextual variables can push the DMUs further away from the frontier or influence the distribution of efficiency scores among peers; however, they cannot change the maximum attainable limit of production or the position of the frontier. Thus, to meaningfully use the SW approach, one has to test the data that describes the PPS for compliance with the separability condition (Daraio, Simar,

and Wilson 2010; 2018). Otherwise, one cannot obtain meaningful statistical inferences in the second stage regression. In fact, a test of the separability condition (Daraio, Simar, and Wilson 2010, page 23) concluded that the empirical example in the original SW study gave "results that are meaningless". A recent study proposes a test of the separability condition that relies on the central limit theorem (Daraio, Simar, and Wilson 2018).

For the PPSs that violate the separability condition, Simar and Wilson suggests (2011; 2015) that the "safest" way handle the contextual variables is to estimate the conditional efficiency measure (Badin, Simar, and Daraio 2012). In their follow-up publication they re-iterate this claim and argue that if the test of the separability condition fails, even the first stage DEA results could be misleading (Daraio, Simar, and Wilson 2018). Based on this suggestion, we explore conditional efficiency measures as a complementary TSA method. Conditional Efficiency Measures (CEMs), leverage the idea of formulating the position of the frontier from a probabilistic perspective using joint probability distribution functions assuming that the input-output variables are jointly distributed with the contextual variables (Cazals, Florens, and Simar 2002). CEMs usually relax the convexity assumption by assuming a free disposable hull (Deprins et.al 1984, 2006) and focuses on understanding the influence of outliers on the frontier by using approaches such as the order-m (Daraio and Simar 2005) or order-α (Aragon et.al 2005). However, the conditional measures are subject to issues that originate from the selection of the bandwidth that determines the admissible set of peers. Data based approaches have been proposed to address this problem (Daraio, Simar, and Wilson 2010). Badin et.al. (2012; 2014) propose an extension of the order-m model with a second stage regression, that allows to differentiate between the impact of contextual factors on the frontier and the individual observations. Conditional order-m approaches have been with extended with kernel regression frameworks to investigate contextual influences in a wide array of application areas (Verschelde and Rogge 2012; Dewitte et al. 2020).

4.2.3 Single Stage Estimation Approaches

In this subsection we will briefly cover the methods that pursue the single stage estimation of the influence of contextual variables, that are led by the semi-nonparametric approach (JK) (Johnson and Kuosmanen 2011). Similar to SW, JK was also motivated by

Banker and Natarajan's (2008) two stage approach. JK argued that two-stage regression methods suffer from the finite sample bias of DEA that carries over from the first stage. Consequently, the coefficients of the second stage regression are biased, especially when the input variables are correlated with the contextual variables. To resolve this issue, JK introduced a convex nonparametric and non-linear programming formulation (Gstach 1998; Groeneboom, Jongbloed, and Wellner 2001). This transformation is mathematically valid as it was previously shown that DEA could be formulated as a constrained convex nonparametric least squares regression (Kuosmanen 2008; Kuosmanen and Johnson 2009). Consequently, the frontier is computed in a single-stage nonparametric way and the influence of contextual variables are decomposed as inefficiency terms similar to the parametric stochastic frontier approaches. It is important to note here that, identical to the SW's interpretation of the contextual factors, JK assumes that the contextual variables are mechanisms that reduce a DMU's efficiency, and they do not alter the position of the frontier (Johnson and Kuosmanen 2011, 221). This is a strong point of departure from TSS's interpretation of the contextual variables, since from the TSS perspective, these factors shape the PPS and therefore have a direct implication on the position of the frontier.

Certain properties of JK render it more flexible compared to SW. First, JK can handle both categorical and ordinal contextual variables. Additionally, it is not bounded by any sign or correlation restrictions and remains consistent even when the noise term is unbounded. Besides, JK is normally distributed and it asymptotically converges. Thus similar to SW, it allows the use of statistical testing for asymptotic inference. Although the empirical usefulness of JK is documented when the approach is adopted by the Finnish electricity regulation authority(Kuosmanen and Kortelainen 2012), the verification in the paper is based on simulated data that relies on a pre-specified functional form to generate the DMUs. Case studies that investigate influence of contextual variables on agriculture production have been published, documenting the JKs' interpretation of these variables (Vidoli and Ferrara 2015). However, a strict limitation of this approach is that it is univariate output, meaning that it only allows for a single output variable. Given that many complex processes such as INFRABEL TCCs produce multiple outputs, JK's univariate output structure renders its use infeasible in complex system problems without an extension

to a multi-output case. Therefore, we skip the implementation of this approach for the scope of this study.

4.3 Methodology

In this section we will present our dataset, discuss its characteristics and our proposed strategy for dealing with data limitation issues. We then provide an overview of discussed analytical approaches.

4.3.1 The Data

This study is enabled by a custom generated INFRABEL dataset that is obtained from nine different TCCS for a random weekday from May, 2019. The data is rich and highly granular as all observations are recorded by the same business intelligence tool that captures all Controller operational decisions. The universe dataset contains 1,478 observations that correspond to one hour of traffic control activities that we treat as the time frame of our DMUs. We consider one hour of Controller activities as an appropriate time-frame to capture sudden variations in the contextual conditions that could originate from the state of railroad. INFABEL traffic is highly dynamic and varies considerably with contextual factors that originate from the demands of the community (e.g., rush hours), the physiological state of the Controller (e.g., mental fatigue), and the resulting variations in the complexity of the decisions that need to be made. We provide variable definitions in Table 4-1.

Table 4-1 The Data and Variable Definitions

Variable Name	Description
Controller Skill Days	Number days the Controller was assigned for the designated role.
Movement Decisions	Time spent (in seconds) for opening signals by Controllers.
Adaptation Decisions	Time spent (in seconds) for decisions that change the state of the railroad such as merging or splitting trains, re-routing of trains, or special procedures at single-track lines. Performed manually by Controllers.
Anticipation Usage	Measure (in seconds) of Controller time spent using the forecast tool that anticipates the future state of the network.
Responded Phone Calls	Number of phone calls addressed by Controllers. Controllers routinely receive phone calls from other INFRABEL personnel about decisions that require further information.
Traffic Complexity	Measure of traffic complexity of control area. Estimated by using the number of control signal passes and the performed adaptation decisions.

Traffic Density	Measure of traffic density in the control area. Calculated by dividing the number of train movements with the number of large traffic control signals controlled by that Controller.
Fatigue Level	Mental fatigue of Controllers. It is calculated by INFRABEL's predictive tool that is conceptually based on the fatigue Risk Index (Roets and Christiaens 2017; Folkard, Robertson, and Spencer 2007).
Delay	Average train delays within the control area. Measured from the scheduled time in seconds. Large delays are due to freight trains. Trains running before schedule (negative delays) can also perturb traffic flows, and therefore are considered through their absolute value.

One of the motivations of this paper is to document how the respective insights of the two approaches vary with respect to data availability. After all, many DEA application papers lack access to organizations and consequently to rich datasets that holistically represent the transformation process. Similarly, unlike the simulated experiments, collecting data from complex processes to implement a DEA approach is an exhaustive task that is subject to many pitfalls such as measurement errors, organizational access issues, and other resource limitations. Besides, it may not be possible to measure or capture all the important sources of information that are necessary to properly abstract a transformation process from a DEA perspective. To replicate this highly likely scenario and to be able to compare the performance of each method in the presence of data limitations, we purposefully subset our dataset, and create an artificially restricted dataset that only includes observations from as single TCC. Notice that we preserve the depth of analysis by keeping the same number of variables for the limited data set. Similarly, we preserve the same time horizon for both datasets to avoid issues that could originate from auto correlation over time. While we recognize that a adopting a longer time horizon would have provided interesting insights, we purposefully chose not to do so to avoid additional variation issues that could originate from the day of the week, autocorrelation among days, among others.

Our justification for creating this artificially limited subset is that, many large-scale application studies require installation of a measurement device, which is usually done for a pilot DMU. Hence, it is quite likely that an analyst might have access to data, however, it could be from a single pilot or experimental source. We restricted our dataset to demonstrate this scenario. Another important aspect that we purposefully leave out of the scope for this study is the dimension limitations. To elaborate, not all the input, output, and contextual variables that are considered necessary to holistically model the production

process may not be available to the analyst. While we realize that many DEA applications could be limited from this perspective, we consider it as a complicated problem that needs to be investigated in depth and we leave it for future work. To summarize, the artificially restricted subset in this study includes 173 DMUs from a single TCC and it is large enough to satisfy the minimum number of DMUs required by the heuristics of model formulation (Dyson et al. 2001; Peyrache, Rose, and Sicilia 2019). In Table 4-2, we document the descriptive statistics of the artificially restricted and the universe. Table 4-2 indicates how data limitations could alter the interpretation of the transformative process, most clearly demonstrated by the increase in the variable ranges.

Table 4-2 Descriptive Statistics of the Datasets

	Subset *(n = 173)*			Universe (*n = 1,478*)		
Variable Name	**Mean**	**Range**	**Std. Dev**	**Mean**	**Range**	**Std. Dev**
Controller Skill Days	2,107.55	[204; 2,989]	922.66	1,881.2	[204.0 ; 2,997.0]	875.82
Movement Decisions	52.34	[0 ; 117]	34.51	47.51	[0 ; 175]	32.3
Adaptation Decisions	237.16	[0 ; 823]	160.28	219.7	[0 ; 1,416.0]	181.34
Anticipation Usage	2.66	[0 ; 70]	9.18	6.09	[0 ; 230]	18.43
Responded Phone Calls	1.49	[0 ; 12]	2.27	1.49	[0 ; 16]	2.24
Traffic Complexity	1,891.37	[0 ; 27,000]	3,454.42	1,029.9	[0 ; 27,000]	1,547.1
Traffic Density	351.16	[0 ; 882.88]	248.08	355.49	[0 ; 1,356.16]	267.84
Fatigue Level	1.25	[1.02 ; 1.79]	0.15	1.24	[1.02 ; 2.35]	0.17
Delay	368.13	[0 ; 3,447.5]	469.35	475.75	[0 ; 22,083]	975.0

To discuss the characteristics of our datasets, we provide Table 4-3 and Table 4-4 that describe the Pearson correlations among variables. We observe low correlation among the contextual variables in both datasets, with a maximum correlation of -0.491 between traffic density and complexity. This indicates that our selection of the factors were appropriate in terms of capturing orthogonal influences that are hard to describe by the other variables. We also observe that increasing the sample size reduces the correlation among variables. Increasing the sample size leads to sign changes among certain correlations, such as between z1-y3 (traffic complexity and anticipation usage) and x1-y2 (controller skill days adaptation decisions). This demonstrates the importance of the sample size in terms of explaining the behavior of complex production processes. By selecting both datasets from a single day, we prevent additional unobserved behaviors from being introduced into our

dataset, e.g. autocorrelation between consecutive days, weekends, scheduled maintenance activities, holidays etc.

Table 4-3 Pearson Correlation Matrix for the Subset

Variable Name	x1	y1	y2	y3	y4	z1	z2	z3	z4
x1 Controller Skill Days	1.000								
y1 Movement Decisions	-0.139	1.000							
y2 Adaptation Decisions	0.256	0.398	1.000						
y3 Anticipation Usage	-0.145	0.003	0.167	1.000					
y4 Responded Phone Calls	0.250	0.275	0.628	0.089	1.000				
z1 Traffic Complexity	0.034	-0.504	0.024	0.210	-0.131	1.000			
z2 Traffic Density	-0.347	0.898	0.195	-0.027	0.107	-0.491	1.000		
z3 Fatigue Level	-0.136	-0.379	-0.283	-0.088	-0.283	0.330	-0.247	1.000	
z4 Delay	0.093	-0.389	-0.162	-0.048	-0.035	0.172	-0.388	0.206	1.000

Table 4-4 Pearson Correlation Matrix for the Universe

Variable Name	x1	y1	y2	y3	y4	z1	z2	z3	z4
x1 Controller Skill Days	1.000								
y1 Movement Decisions	-0.022	1.000							
y2 Adaptation Decisions	-0.007	0.327	1.000						
y3 Anticipation Usage	-0.153	0.295	0.140	1.000					
y4 Responded Phone Calls	0.072	0.103	0.321	0.036	1.000				
z1 Traffic Complexity	-8.60e-05	-0.291	0.280	-0.004	0.041	1.000			
z2 Traffic Density	-0.079	0.898	0.236	0.278	0.046	-0.284	1.000		
z3 Fatigue Level	-0.097	-0.168	-0.093	-0.064	-0.044	0.123	-0.145	1.000	
z4 Delay	-0.029	-0.183	0.004	-0.064	0.108	0.099	-0.189	0.0865	1.000

4.3.2 Model Formulation

We use the microeconomic production theory to map the variables in our dataset to DEA variables and provide the blackbox diagram of the model in Figure 4-1. In the introduction, we mentioned that we aim to preserve construct validity by extending our previously verified model formulation (Topcu, Triantis, and Roets 2019). We expand our previous output oriented variable returns to scale model (Banker et.al 1984) with additional sources of information that have recently become available, which are: anticipation usage, responded phone calls, and delays. The fundamental assumption of this representation of traffic control activities is that, given a Controller with X level of experience can handle Y

amount of tasks under Z circumstances, then a more experienced controller should be able to handle larger amount of tasks in a given time. To elaborate, the only resource that is consumed by the traffic control process is the experience level of the Controller since that employee cannot be utilized for any other activities for the duration of that time. The output variables represent the routine traffic control tasks, as defined in Table 4-1. Finally, the contextual variables represent the "quasi factors" that describe the environment in which the transformative process takes place. These variables are uncontrollable by the DMU, are not produced nor consumed by the transformative process, yet have a strong influence on the process. A detailed discussion of the omitted variables, why these variables were considered appropriate, and how the mapping to the DEA variables is articulated is provided elsewhere (Topcu, Triantis, and Roets 2019). We proceed with the methodology of each investigated algorithm.

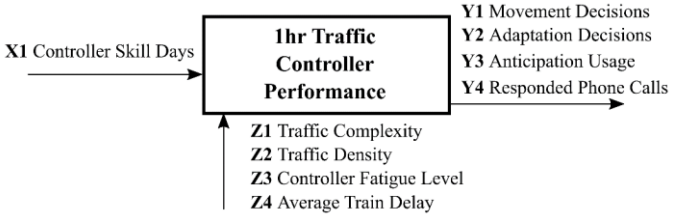

Figure 4-1 Controller DEA Black-box with Contextual Variables

4.3.3 The Robust Multivariate Method (TSS)

The theoretical motivation of the TSS approach is to preserve the homogeneity (or comparability) assumption via clustering, it is applicable as long as the contextual variables are continuous, and it achieves this by integrating three distinct data-analytics algorithms. We provide a flowchart of the TSS approach in Figure 4-2 and proceed to explain the details of the algorithm. Once the dataset is assembled, the first step of the TSS is the test of the PPS for homogeneity. This is conducted through the ROBPCA analysis (Hubert, Rousseeuw, and Branden 2005). ROBPCA reduces the dimensions of the dataset into principal components that describe most of the information contained within the data. The observations are then classified based on their relative position with respect to the axes of principal component hyperplanes, measured by two distances: score (parallel to the hyperplane) and orthogonal (vertical to the hyperplane) distances. Based on the relative distance of the observation with respect to these axes, ROBPCA categorizes observations

into four categories: regular observations that constitute the bulk of the observations, and the orthogonal, good, and bad leverage points, which are outliers compared to the rest of the data. Within these three leverage categories, the most extreme cases are represented by bad outliers, which the TSS approach considers as observations that strictly violate the *homogeneity* or the *comparability assumption*. However, rather than considering these observations as outliers and discarding from the dataset, we purposefully include them in our analysis because (i) our measurement tool is highly reliable therefore it is certain that these events actually occurred and (ii) these events could introduce safety risks for the operation of the STS therefore require consideration.

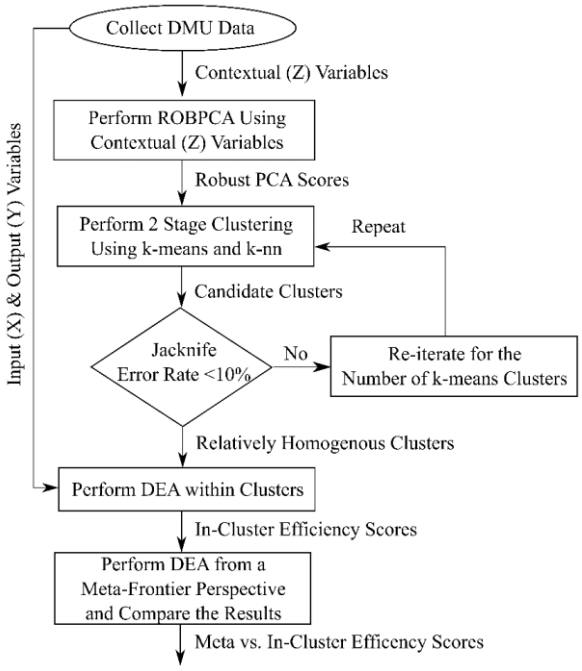

Figure 4-2 Flowchart for the TSS Algorithm

The second step of the TSS algorithm is to use the robust principal component scores to formulate relatively homogenous clusters, which is achieved through the two stage clustering that integrates the nearest neighbor clustering (Wong and Lane 1983) with k-means (Hartigan and Wong 1979). This is an iterative process where each resulting candidate clusters are evaluated for their jackknife error rates (Miller 1974) until a candidate provides less than 10% error. In the case that more than two candidate clusters provide less than 10%, the candidate with minimum error rate is preferred. At the end of this process the relatively homogenous clusters are obtained, which are then evaluated from

both an in-cluster perspective and with respect to the Meta Frontier that disregards the influence of contextual factors. Any nonparametric efficiency measurement algorithm can be utilized for this step, as long as the same algorithm is used to compute the position of the in-cluster and Meta Frontier. Consequently, the aggregate impact of contextual factors are quantified through the technology gap (Battese, Rao, and O'Donnell 2004).

4.3.4 The 2-Stage Approaches

Based on our literature review, TSA approach to handle heterogeneous PPSs is visualized in Figure 4-3. The first step is to approach the problem from a SW perspective and check for the separability condition. This is important because if the separability condition does not hold, it means that the boundary of maximum attainable performance is dependent on the contextual conditions Z and the second stage regression should be based on the conditional frontier (Daraio, Simar, and Wilson 2018) instead of the SW algorithm (Simar and Wilson 2007). One way of testing the separability condition (Daraio, Simar, and Wilson 2018) is through the comparison of frontiers for the conditional (e.g. order-m) (Cazals, Florens, and Simar 2002) and unconditional cases. This is based on the assumption that, if a PPS Ψ satisfies the separability condition, then all of its conditional interpretations PPSs Ψ^c should display similar characteristics. Therefore, the sample mean of unconditional efficiency scores $\widehat{\mu_n}$ computed by disregarding the contextual variables (Z_i) and only using the input (X_i) and output variables (Y_i) is shown in Equation 1 and it should be close to the mean of conditional efficiency scores $\widehat{\mu_{c,n}}$ computed through Cazals et.al. 2002 as shown in Equation 2. If the PPS violates the separability condition, the difference between the two means represented by Equation 3 would be considerably larger than zero.

$$\widehat{\mu_n} = E\left[\sum_{i=1}^{n} \hat{\theta}\,(X_i, Y_i|\, S_n)\right] \quad (1)$$

$$\widehat{\mu_{c,n}} = E\left[\sum_{i=1}^{n} \hat{\theta}\,(X_i, Y_i|Z_i,\ S_n)\right] \quad (2)$$

$$\varepsilon = \widehat{\mu_n} - \widehat{\mu_{c,n}} \quad (3)$$

Once the separability condition is tested using Equations 1-3, one could proceed to the next stages following one of the two recommended paths visualized in Figure 4-3. If the separability condition indeed holds and the value of ε described in Equation 3 is sufficiently small than one could proceed with the traditional SW algorithm (Simar and Wilson 2007).

If the separability condition does not hold, then the recommended route is to compute a conditional efficiency measure, that could follow the order-m approach (Cazals, Florens, and Simar 2002; Daraio and Simar 2005; Badin, Simar, and Daraio 2012).

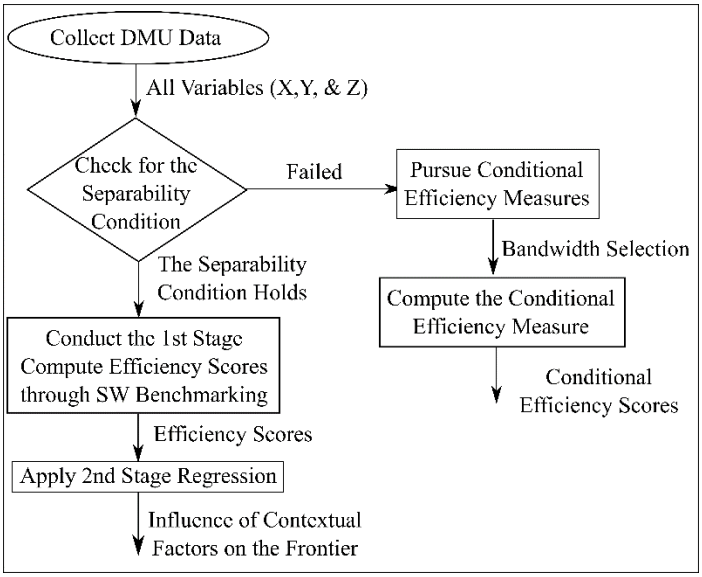

Figure 4-3 the 2-Stage Approach (SW on Left, Cazals on right)

4.4 Results & Discussion

4.4.1 Complementary Roles

In this subsection, we will cover the complementary roles of the efficiency measurement methods under contextual heterogeneity. As previously discussed, we will explore the sensitivity of insights to sample size by presenting our analysis using two datasets, a universe that is composed of observations from 9 TCCs and a subset that only contains information from a single TCC. We will discuss the complementary roles in the following order. First, following a TSS strategy, we will investigate the fraction of influential observations in both PPSs through ROBPCA,. In terms of exploring the complementary roles for this study, we use ROPBCA as a proxy test for the homogeneity assumption. If we observe a substantial fraction of influential observations, we consider that as an indicator of heterogeneity. Following TSS approach, we then use the results of the ROPBCA analysis to formulate relatively homogenous clusters in which DMUs could be evaluated by considering the aggregate impact of contextual influences. We then proceed to the TSA, and provide a test of the separability condition following the rule of

thumb test (Daraio, Simar, and Wilson 2018). We discuss the results based on feedback from domain experts and explain the contribution of contextual factors by using a two-stage method.

4.4.1.1 TSS - Test of the Homogeneity Assumption through ROBPCA

We use the contextual variables for both the subset and the universe PPSs and conduct the ROBPCA analysis (Hubert, Rousseeuw, and Branden 2005). For the subset we observe that a single principal component yields the minimum standard deviation with 596.1, where for the universe, two principal components yield minimum standard deviation of 927.38. In the case of the universe, the first PC explains 88% of the variability and it is significantly influenced by the traffic complexity and density, followed by delays. This indicates that the relative variance of the fatigue level is low, which could be attributed to INFRABEL's operational practice of considering fatigue levels in staff scheduling. We observe by looking at the variable loadings that traffic complexity and traffic density are the leading factors in terms of determining the shape of the principal component axes. We present the details in Table 4-5.

Table 4-5 Loadings of Principal Component Axes

	Subset n = 173, Single TCC	**Universe** n = 1,478, All 9 TCCs	
Principal Components	**PC1**	**PC1**	**PC2**
Standard deviation	596.8	680.347	247.040
Proportion of Variance	1	%88.3	%11.7
Variable Loadings			
z1 Traffic Complexity	0.988	-0.990	-0.113
z2 Traffic Density	-0.143	0.116	-0.992
z3 Fatigue Level	8.95e-06	-1.097e-05	5.77e-05
z4 Delay	0.048	-0.073	-0.036

We present the plot of the outlier identification in Figure 4-4. In Figure 4-4, the left side represents the subset and the right side represents the universe PPSs. The vertical lines represent the cutoff distances, where the score cutoff values are [2.24; 2.71] and the orthogonal cutoff values are computed as [644.19; 488.10] for the subset and the universe

PPS respectively. The subset has 13 bad leverage points that represent extreme outliers in the dataset where as the universe has 39, interestingly the subset includes almost twice the fraction of bad leverage points. We attribute this to the fact that our subset is the busiest Traffic Control Center (TCC) that controls the densest area in the entire network. Similarly, the fraction of good leverage points that represent the second largest group of influential observations, represented with green in Figure 4-4, are almost threefold in the subset. Consequently, the universe is composed of 76% regular observations whereas the subset only includes 69%. From this perspective, it is safe to argue that the subset PPS is more heterogeneous than the universe. However, both PPSs demonstrate heterogeneous characteristics. From a TSS perspective, this requires further classification. We will investigate how this observed heterogeneity relates to the separability condition in the following subsections.

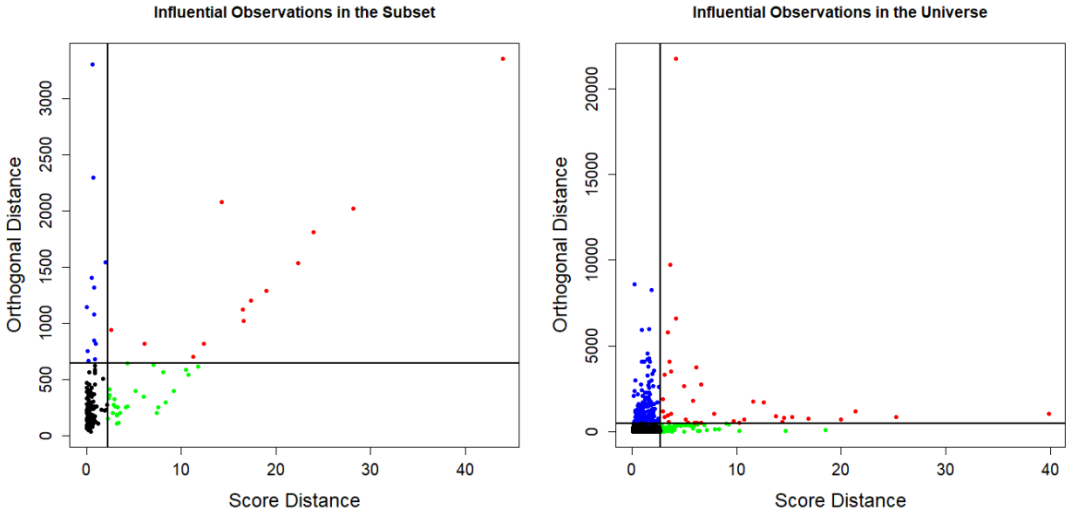

Figure 4-4 Test of the Homogeneity Assumption through ROBPCA

4.4.1.2 Formulation of Relatively Homogenous Clusters

The TSS approach uses the ROBPCA (score and orthogonal) scores that represent the distance of observations from the principal axes, to inform the formulation of relatively homogeneous clusters. The clustering in TCC is performed in two stages. First, a k-nn nearest neighbor clustering is performed with usually two or three nearest neighbors. In the second stage, the resulting nearest neighbor clusters are used to inform the search for the ideal number of k-means clusters that would represent the dataset. We compare the

jackknife error rates for the k-means clusters and pick the number of k-means clusters that has the minimum jackknife error, which we assume to provide an accurate statistical representation of the data.

For the case of the subset, four clusters appear to be the ideal number as increasing the number of clusters from 3 to 4 reduces the within cluster sum of squares by 40%, yet an additional increase from 4 to 5 returns a much lower decrease of only 20%. Similarly, the universe set is accurately explained by four clusters. Figure 4-5 presents cluster distributions, visualizing the scatter of homogenous performance groups around the principal component axes.

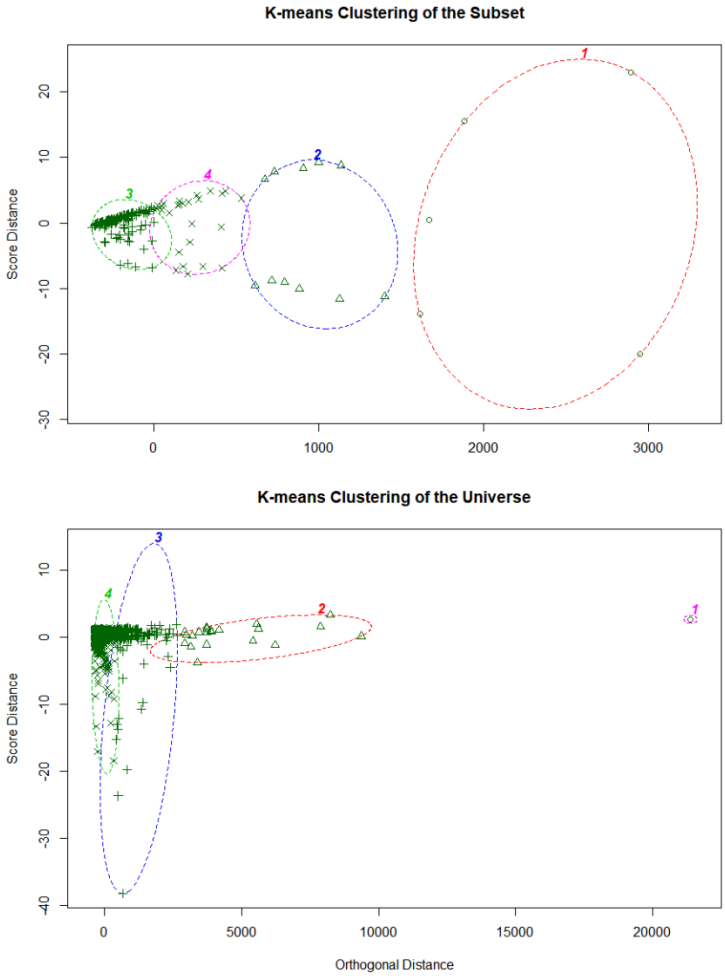

Figure 4-5 K-Means Clusters Around the Principal Component Axes

We provide Table 4-6 to elaborate on the cluster characteristics. As indicated by Figure 4-5, the first cluster for both sets are scarcely populated, indicating that these observations

are located farthest away from the rest as measured by ROBPCA. For case of the subset, we observe that extreme observations experience significantly higher traffic complexity, delays, and fatigue levels. Similarly, the first cluster in the universe exhibits considerably higher delays; in fact, it's only populated by a single observation. By isolating these observations from others, TSS ensures that these cases are evaluated by considering their unique characteristics. For instance, the most densely populated cluster in each dataset has the lowest average traffic complexity. Given that these regular observations constitute almost 90% percent of the universe, the uniqueness of extreme observations would have been overlooked if the problem was approached from a meta frontier perspective. In other words, each cluster in this multivariate perspective represents an independent frontier, each providing a different mechanism for the same production technology to be efficient. Identification and classification of these subsets produce managerial utility by documenting how frequently these circumstances occur and how they differ from the usual operation. We continue our comprehensive analysis with the TSA.

Table 4-6 Relatively Homogenous Cluster Characteristics – Mean Values

	Subset n = 173, Single TCC				**Universe** n = 1,478, All 9 TCCs			
	CL1 n = 5	**CL2** n = 11	**CL3** n = 131	**CL4** n = 26	**CL1** n = 1	**CL2** n = 21	**CL3** n = 148	**CL4** n =1,308
Z1 Traffic Complexity	10,700	6,914	1,025.56	2,434.62	1,409.09	1,668.62	1,735.80	939.58
Z2 Traffic Density	6.73	1.67	426.12	153.72	90.90	151.46	187.27	378.00
Z3 Fatigue Level	1.48	1.36	1.21	1.31	1.24	1.32	1.29	1.23
Z4 Delay	2,045.29	862.27	219.83	583.76	22,083.52	5,290.88	1,581.31	256.83

4.4.1.3 Test of the Separability Condition

As the first step of the TSA approach, we check for the separability condition using Equations 1-3. We prioritize this step before proceeding into details because we would like to document the potential issues once could face when this test is ignored. Moreover, the test informs the following steps in terms of interpreting the DMUs. For the test, we use the

rule of thumb and compare the unconditional efficiency scores. We utilize the "Benchmarking" (Bogetoft and Otto 2010) and "rDEA" (Besstremyannaya 2011) packages and present the results in Table 7.

Table 4-7 Conditional vs. Unconditional Efficiency Scores

	Subset n = 173, Single TCC	**Universe** n = 1,478, All 9 TCCs
Mean Meta Frontier Efficiency Scores (Banker et.al 1984)	0.592	0.398
Mean Bias-Corrected Efficiency Scores (Simar and Wilson 2007)	0.547	0.373
Mean Order-m Conditional Efficiency Scores (Cazals et.al. 2002)	0.713	0.634

For the separability condition to hold, the unconditional efficiency measures given in the first row have to be equal or similar to the conditional efficiency scores provided in the third row. Clearly, the difference is larger than zero. More interestingly, the difference between the two values increase with sample size. Therefore, we reject the separability condition. However, we proceed with our analysis as if there are no issues with the separability condition, so that we can demonstrate the potential consequences of doing so.

4.4.1.4 SW – Influence of Contextual Variables through 2nd Stage Regression

Similar to other two stage approaches, SW relies on a second stage bootstrap based regression to compute the influence of contextual variables on the efficiency scores. From this perspective, in our case the regression results represent the influence of the contextual variables on Controller task workload. Although we rejected the separability condition, we compute the regression results following SW and present the results in Table 4-8. We start our comparison with the intercept, and observe that its variation is so high, that it changes sign within the confidence interval. Given the separability condition was rejected, we expect this behavior and hope that it precisely documents potential issues one could experience if the validity of fundamental assumptions are ignored. A comparison of the subset to the universe reveals that similar sign flips occurs for the influence of fatigue levels. Especially in the case of fatigue levels, we observe the variation to be considerable.

The standard deviation of errors decrease with increasing sample size, indicating that accuracy of the SW increases with new information.

Table 4-8 Influence of Environmental Variables – Simar Wilson 2007

	Subset n = 173, Single TCC, α = 0.05			**Universe** n = 1,478, All 9 TCCs, α = 0.05		
	$\hat{\hat{\beta}}$	**Confidence Interval Low 2.5%**	**Confidence Interval High 97.5%**	$\hat{\hat{\beta}}$	**Confidence Interval Low 2.5%**	**Confidence Interval High 97.5%**
Constant	25.636	-2.943	51.575	10.986	-19.177	37.238
Z1 Traffic Complexity	-0.120	-0.300	-0.053	-0.083	-0.153	-0.044
Z2 Traffic Density	-16.104	-39.042	-7.516	-8.046	-14.497	-4.395
Z3 Fatigue Level	42.269	-6.481	90.165	17.255	-14.176	43.552
Z4 Delay	-0.071	-0.253	-0.009	0.018	-0.003	0.040
$\hat{\sigma}$ **(std dev. of errors)**	67.828	38.471	115.365	72.216	52.736	98.817

For the case of the subset, traffic complexity, density, and delays are interpreted as negative influences where the influence of fatigue appears positive yet ambiguous due the confidence interval. A similar tendency is observed in the universe. In order to check whether these insights contradict reality or not, we analyze the results with INFRABEL experts. Our discussion reveals that Controllers consider complexity and density as the leading factors that increase their workload, whereas the influence of fatigue is considered negligible. This is exactly the opposite of SW results. We would like to emphasize that, the SW approach works fine within the axiomatic boundaries that are defined with the separability condition. However, the separability condition simply does not hold for this specific application. For the DEA community, this highlights the importance of establishing some form of a verification mechanism. If we didn't have access to the organization and were interpreting these results without validating or testing the separability condition, we would intuitively conclude that fatigue levels considerably increase the workload. However, as indicated by this simple face-validation activity, the mitigating influence of contextual factors may not necessarily have intuitive or expected consequences, due to the complexity of the process. We proceed to discuss the efficiency scores.

4.4.2 Interpretation of the Efficiency Scores from Contrasting Perspectives

For both datasets, we compute the efficiency scores following the TSS and TSA approaches and present the results in Table 4-9. The first column in Table 4-10 represent the efficiency scores computed through the traditional BCC algorithm (Banker et.al. 1984), which represent meta efficiency scores and serve as the baseline for our comparison. To recall, SW uses these scores in the first stage to relate to the contextual variables, and computes bias corrected efficiency scores (Simar and Wilson 2007), which we present in the second column. The third column represents the in-cluster efficiency scores computed through TSS (Triantis et.al. 2010). Finally, the fourth column represent order-m efficiency scores, computed for the meta frontier (Cazals et.al. 2002).

Table 4-9 Distribution of Efficiency Scores

	Subset n = 173, Single TCC				**Universe** n = 1,478, All 9 TCCs			
Efficiency Distributions	**BCC Meta Eff**	**SW**	**TSS In-Cluster**	**Order-m**	**BCC Meta Eff**	**SW**	**TSS In-Cluster**	**Order-m**
Mean	0.592	0.547	0.655	0.713	0.398	0.371	0.419	0.634
Median	0.592	0.562	0.662	0.738	0.362	0.346	0.376	0.574
Standard Deviation	0.284	0.258	0.270	0.410	0.238	0.215	0.248	0.494
Number of Efficient DMUs	17	0	30	40 >1	31	0	51	289>1

For both datasets, SW scores are lower than Meta efficiency scores while TSS and order-m scores are higher. It is expected for TSS results to be higher than the Meta efficiency scores, especially given that certain clusters are relatively small in size due the unique contextual conditions they face, and consequently, these observations experience the greatest increase in their technology gap. We observe that increasing difficulty of operational conditions lead the SW algorithm to penalize these observations and reduce their efficiency score while TSS rewards them with higher efficiency scores. This contrast is also observed in the number of fully efficient observations, as TSS approach allows each specific cluster to have its own efficient units. We observe that, as expected, the average

bias correction represented by the gap between SW and Meta efficiency scores, decreases with increasing sample size. On average, TSS scores are 10% higher than meta efficiency scores for the subset, and this difference is observed to decrease with increasing sample size. Combining this information with the cluster characteristics, we could be argue that the technology gap between efficiency scores in TSS increase when the sample size is small. An important distinction we would like to make at this point is that, since order-m scores are not bounded between 0-1, their mean could be misleading.

In order to further investigate, how these different perspectives would interpret the daily operation of a DMU, we provide Figure 4-6. Figure 4-6 visualizes the hourly assessment of an anonymized control zone that is operated by three rotating Controllers through the day. We specifically choose a workstation that experiences three distinct operational modes as identified by the clustering analysis, so that we can focus on the extreme cases. In Figure 4-6, color codes are like the following. Blue represents Meta efficiency scores, red SW, yellow TSS, and purple for conditional efficiency scores.

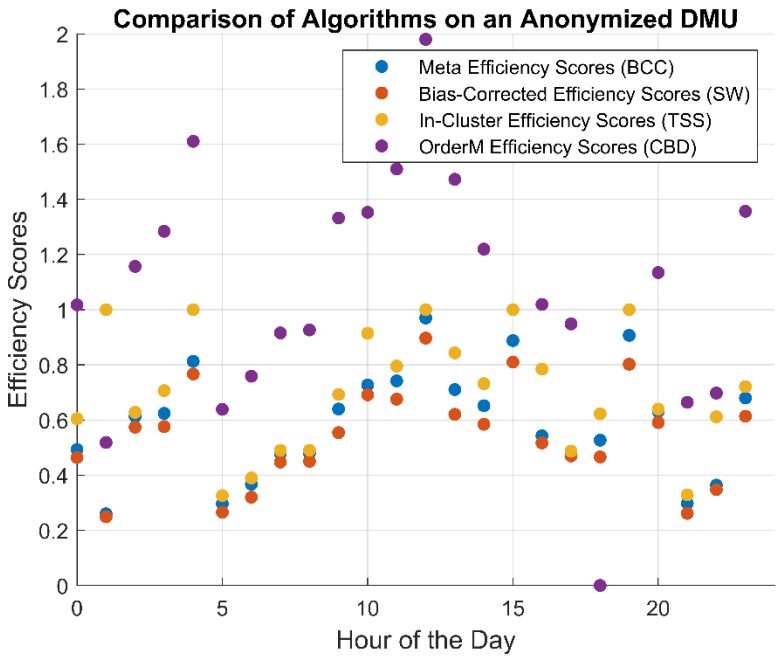

Figure 4-6 Comparison of Algorithms on an Anonymized DMU

We notice that conditional efficiency scores demonstrate different characteristics compared to other three algorithms. Recall that order-m computes efficiency scores based

on the behavior of the frontier by adding and subtracting DMUs to the reference PPS. When a DMU has a high efficiency score, indicated by the agreement of other three algorithms such as in the case of the 12PM in Figure 4-6, we observe that the conditional efficiency measure consistently overshoots. Interestingly, it also undershoots, such as in the case of 6PM observation that has a conditional efficiency score of zero. We attribute this peculiar behavior to the high fraction of outliers in the data as previously indicated by the ROBPCA. Consequently, this begs us to question the common suggestion in the literature that proposes to pursue a conditional efficiency measure when the separability condition is not satisfied. Our observations indicate the exact opposite of this suggestion. Clearly, conditional efficiency measures could mislead for heterogeneous datasets such as this one where the separability condition does not hold. This also demonstrates the importance of field testing algorithms in addition to the artificially populated statistical lab tests.

For the cases where the conditional efficiency scores far exceeded the values of one, for example 5AM and 10 AM, we observe that TSS scores are consistently indicating higher scores, which are created by the differences in the contextual factors of their respective clusters. These observations are usually located in different clusters from the rest of the observations. From this perspective, it appears as TSS exhibits a similar reaction to extreme observations; however, the response is more robust or controlled. When the observations are not in a different cluster, such as in the case of 8 PM, TSS scores are closer to the Meta scores. Focusing on the interpretation of other algorithms, we observe that SW and Meta scores are considerably similar. While the difference between SW and Meta efficiency scores vary from 0 to 30% of the meta scores, median SW score is 4.5% less than the Meta score. Driven by the structure of their distinct algorithms, TSS and SW differ in terms of reflecting contextual influences on the computed efficiency scores. We observe that, TSS compensates observations that perform in extreme operational conditions by restricting their set of peers through clustering, which in return yields higher scores; while SW reduces the efficiency scores of these observations as part of its sample bias-correction efforts.

4.4.3 A Taxonomy of Synthesis

We explored the complementary and contrasting roles of performance measurement approaches that are concerned with contextual variables. Our comprehensive discussion focused on their fundamental understanding, mathematical structure, application areas, limitations, and practical utility. We provide a synthesis of our discussion in Table 4-10 as a structured taxonomy.

Table 4-10 Comparison of TSS and TSA

	The Multivariate Method (TSS)	The 2-Stage Approaches (TSA)
Assumptions Regarding Heterogeneity & Contextual Variables	• Contextual (Z) variables influence the PPS both in terms of the position of the frontier and the efficiency scores. • Aims to preserve homogeneity through robust multivariate statistics.	• SW assumes Contextual (Z) variables do not influence the position of the frontier but influence the efficiency scores. • SW assumes that the production technology is separable. • Order-m does not require separability and assumes the contextual variables influence the frontier and the individual scores.
Computation Time	• Short computation time, no Monte Carlo Iterator.	• SW has relatively long computation times due to the Monte Carlo iterator, especially when n is large. • With a tendency to larger and larger datasets (or even "big data"), this issue becomes more prominent and can even impede real-world implementation. • Order-m has similar properties.
Robustness & Bias Effect	• PPS is parsed into relatively homogenous clusters through robust statistics. • Since the computation of the in-cluster, efficiency scores do not include DMUs that operate in favorable conditions; the bias of extreme observations on the frontier are implicitly reduced compared to the Meta frontier perspective.	• SW Bootstrapping eliminates sample bias through Monte-Carlo iterations. • If the fraction of influential observations is high, order-m might overshoot efficiency estimations, especially with for high efficiency observations.
Flexibility & SW Availability	• Facilitates all DEA models including non-discretionary variables, returns to scale assumptions, non-convexity etc. • Its modularity allows for integration with other complementary techniques, such as classification or prediction algorithms. • Does not have published unified software, however the steps are available in different formats. ROBPCA is available in R, whereas 2stage Clustering is available in SAS.	• SW has a published R module (Besstremyannaya 2011). • The FEAR package (Wilson) provide the necessary modules for SW, however the combination of these modules in the full algorithm is left to the developer. • There are also published conditional order-m models, however multivariate contextual case is unpublished.
Sample size effect	• Sample size influences the number and characteristics of the clusters. • Sample size have no significant effect on the validity of results.	• Increasing the sample size reduces the standard deviation for SW. • For heterogeneous datasets such as this one investigated here, we observe that introduction of additional variability could increase standard deviation for order-m.

Managerial Information	• Allows to identify statistically different performance groups that could represent contextual differences faced during operation. • The first step (robust clustering) provides useful information regarding the characteristics and frequency of extreme operational cases. • Is easier to explain to non-experts.	• SW statistically explains the individual influence of contextual variables and allows for statistical inference tests. • Also provides confidence intervals for the influences. • Order-m methods could be extended with kernel regression techniques to explain individual contribution.
Limitations	• Does not explain the influence of individual contextual variables. • Does not allow for statistical tests. • Since clusters are formed based on robustness, they are formed based on a combination of factors, which may not have a direct practical correspondence. • Handles continuous variables and can handle binary variables to a certain extent.	• SW's separability condition imposes strong restrictions in terms of application areas and it is rarely checked in the literature.

4.5 Discussion, Conclusions, and Future Work

We investigated the complementary and contrasting roles of performance measurement methods that are concerned with contextual influences. We were motivated by the fact that many transformation processes experience high variations in terms of the contextual factors that influence their operations. These influences need to be considered to make effective managerial decisions. In order to document how these contextual factors could shape a complex production process; we adopted an empirical approach and investigated INFRABEL's operational infrastructure control system. The investigated system is of significant complexity and, experiences drastic social and technical contextual variations during operation. We were able to study a rich dataset that allowed analyses on an hourly resolution. This allowed us to explore the utility of the two approaches without the need to rely on restrictive assumptions, unverifiable case studies, or manufactured lab tests. We believe the complexity of the investigated transformation process constitutes an excellent case where one can explore how the collective insights of some of the existing methods could help to inform better decisions. For this purpose, we explored a multivariate strategy and a two-stage strategy, and documented their sensitivity to sample size.

We conducted our analysis in conjunction with INFRABEL domain experts for face validation and organized our results in a structured taxonomy. The fundamental focus of

the multivariate strategy, was to derive relatively homogenous clusters in which the units are comparable. This served as a useful tool for exploring the influential DMUs in the dataset. Identification and classification of clusters produce managerial utility by documenting how frequently these circumstances occur and how they differ from the usual operation. Additionally, use of ROBPCA, or a similar robust influential observation technique, could help managers to identify and classify the aggregate impact of contextual influences on their operations, irrespective of the characteristics of the production technology. Similarly, component loadings in ROBPCA could serve as a surrogate for understanding the relative influence of contextual factors. This could potentially serve as an alternative to the second stage regression in the two-stage approaches.

Going back to the separability condition, while the need for imposing such strong restrictions is explained based on the statistical properties of the SW algorithm, we observed that it does not hold in the case of INFRABEL TCCs. This could have been intuitively expected, since no transformation process can be considered in void, decoupled from its environment. Regardless, we proceeded with the second stage regression to document how the contextual influences would be interpreted by the SW approach, if the separability test was ignored. We observed the regression results to be misleading and contradicted the opinions of INFRABEL experts. Our observations support many others in our research community (Olesen and Petersen 2009; Bădin, Daraio, and Simar 2010; Daraio, Simar, and Wilson 2018; Banker, Natarajan, and Zhang 2019). Moreover, since we observed that the separability condition does not hold, we explored the suggestion of Simar and Wilson and employed a conditional efficiency measure. Our analysis documented that, the order-m measure demonstrates an interesting overshoot-undershoot behavior, especially in the presence of outliers. Multivariate robust clustering results were observed to be reacting to these extreme observations in a similar way; however, the reactions were more robust, since the evaluation is bounded by the restriction of the peer set through clustering.

To conclude, we observe that for the case of TSA, statistical concerns are prioritized over empirical validity of the developed methods as verification activities have been delegated to simulated lab experiments instead of on-site implementation studies (Simar

and Wilson 2015). By definition, theory and application cannot be considered as two mutually exclusive entities (Weick 1995) as any self-proclaimed theory that fails to explain a natural phenomenon or establish the connection with reality is not rigorous (Eisenhardt and Graebner 2007; Corley and Gioia 2011). In other words, a mathematical algorithm that is statistically proven to yield consistent results cannot be considered as a "theoretical approach" unless it explains a certain phenomenon. Consequently, we believe that one cannot argue about the validity of a method without providing the supporting empirical evidence that could be obtained from the stakeholders who own and operate the investigated transformation process.

References

Acemoglu, Daron, and Pascual Restrepo. 2017. "Robots and Jobs: Evidence from US Labor Markets." NBER Working Paper No. W23285. https://ssrn.com/abstract=2941263.

Aragon, Y., A. Daouia, and C. Thomas-Agnan. 2005. "Nonparametric Frontier Estimation: A Conditional Quantile-Based Approach." Econometric Theory 21 (2): 358–89. https://doi.org/10.1017/S0266466605050206.

Athanassopoulos, Antreas, and Konstantions Triantis. 1998. "Assessing Aggregate Cost Efficiency and the Related Policy Implications for Greek Local Municipalities - ProQuest." INFOR: Information Systems and Operational Research 36 (3): 66–83. https://search-proquest-com.ezproxy.lib.vt.edu/docview/228468653/abstract/8DA4FDD9D20D4363PQ/1?accountid=14826.

Bădin, Luiza, Cinzia Daraio, and Léopold Simar. 2010. "Optimal Bandwidth Selection for Conditional Efficiency Measures: A Data-Driven Approach." European Journal of Operational Research 201 (2): 633–40. https://doi.org/10.1016/j.ejor.2009.03.038.

———. 2014. "Explaining Inefficiency in Nonparametric Production Models: The State of the Art." Annals of Operations Research 214 (1): 5–30. https://doi.org/10.1007/s10479-012-1173-7.

Badin, Luiza, Leopold Simar, and Cinzia Daraio. 2012. "How to Measure the Impact of Environmental Factors in a Nonparametric Production Model." European Journal of Operational Research 223 (3): 818–33.

Banker, Rajiv D., Abraham Charnes, and William Wager Cooper. 1984. "Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis." Management Science 30 (9): 1078–1092.

Banker, Rajiv D., and Richard C. Morey. 1986. "The Use of Categorical Variables in Data Envelopment Analysis." Management Science 32 (12): 1613–1627.

Banker, Rajiv D., and Ram Natarajan. 2008. "Evaluating Contextual Variables Affecting Productivity Using Data Envelopment Analysis." Operations Research 56 (1): 48–58. http://www.jstor.org.ezproxy.lib.vt.edu/stable/25147166.

Banker, Rajiv, Ram Natarajan, and Daqun Zhang. 2019. "Two-Stage Estimation of the Impact of Contextual Variables in Stochastic Frontier Production Function Models Using Data Envelopment Analysis: Second Stage OLS versus Bootstrap Approaches." European Journal of Operational Research,

Advances in Data Envelopment Analysis, 278 (2): 368–84. https://doi.org/10.1016/j.ejor.2018.10.050.

Barros, Carlos Pestana, Milton Nektarios, and A. Assaf. 2010. “Efficiency in the Greek Insurance Industry.” European Journal of Operational Research 205 (2): 431–36. https://doi.org/10.1016/j.ejor.2010.01.011.

Battese, George E., D. S. Prasada Rao, and Christopher J. O’Donnell. 2004. “A Metafrontier Production Function for Estimation of Technical Efficiencies and Technology Gaps for Firms Operating Under Different Technologies.” Journal of Productivity Analysis 21 (1): 91–103. https://doi.org/10.1023/B:PROD.0000012454.06094.29.

Besstremyannaya, Galina. 2011. “Managerial Performance and Cost Efficiency of Japanese Local Public Hospitals: A Latent Class Stochastic Frontier Model.” Health Economics 20 (S1): 19–34.

Blank, Jos L. T., and Vivian G. Valdmanis. 2010. “Environmental Factors and Productivity on Dutch Hospitals: A Semi-Parametric Approach.” Health Care Management Science 13 (1): 27–34. https://doi.org/10.1007/s10729-009-9104-0.

Bogetoft, Peter, and Lars Otto. 2010. Benchmarking with DEA, SFA, and R. International Series in Operations Research & Management Science. Springer.

Broniatowski, David A., and Conrad Tucker. 2017. “Assessing Causal Claims about Complex Engineered Systems with Quantitative Data: Internal, External, and Construct Validity.” Systems Engineering 20 (6): 483–96. https://doi.org/10.1002/sys.21414.

Cazals, Catherine, Jean-Pierre Florens, and Leopold Simar. 2002a. “Nonparametric Frontier Estimation: A Robust Approach.” Journal of Econometrics, 06 (1): 1–25. https://doi.org/10.1016/S0304-4076(01)00080-X.

Charnes, Abraham, William W. Cooper, Boaz Golany, Larry Seiford, and J. Stutz. 1985. “Foundations of Data Envelopment Analysis for Pareto-Koopmans Efficient Empirical Production Functions.” Journal of Econometrics 30 (1–2): 91–107.

Charnes, Abraham, William W. Cooper, and Edwardo Rhodes. 1978. “Measuring the Efficiency of Decision Making Units.” European Journal of Operational Research 2 (6): 429–444. http://www.sciencedirect.com/science/article/pii/0377221778901388.

Chilingerian, Jon A., and H. David Sherman. 2004. “Health Care Applications.” In Handbook on Data Envelopment Analysis, 481–537. Springer.

Corley, Kevin G., and Dennis A. Gioia. 2011. “Building Theory about Theory Building: What Constitutes a Theoretical Contribution?” Academy of Management Review 36 (1): 12–32. http://amr.aom.org/content/36/1/12.1.short.

Dai, Xiaofeng, and Timo Kuosmanen. 2014. “Best-Practice Benchmarking Using Clustering Methods: Application to Energy Regulation.” Omega 42 (1): 179–88. https://doi.org/10.1016/j.omega.2013.05.007.

Daraio, Cinzia, and Léopold Simar. 2005. “Introducing Environmental Variables in Nonparametric Frontier Models: A Probabilistic Approach.” Journal of Productivity Analysis 24 (1): 93–121. https://doi.org/10.1007/s11123-005-3042-8.

Daraio, Cinzia, Léopold Simar, and Paul W. Wilson. 2010. “Testing Whether Two-Stage Estimation Is Meaningful in Non-Parametric Models of Production.” ISBA Discussion Paper.

———. 2018. “Central Limit Theorems for Conditional Efficiency Measures and Tests of the ‘Separability’ Condition in Non-Parametric, Two-Stage Models of Production.” The Econometrics Journal 21 (2): 170–91. https://doi.org/10.1111/ectj.12103.

Deprins, Dominique, Léopold Simar, and Henry Tulkens. 2006. “Measuring Labor-Efficiency in Post Offices.” In Public Goods, Environmental Externalities and Fiscal Competition, edited by Parkash Chander, Jacques Drèze, C. Knox Lovell, and Jack Mintz, 285–309. Boston, MA: Springer US. https://doi.org/10.1007/978-0-387-25534-7_16.

Dewitte, Ruben, Michel Dumont, Bruno Merlevede, Glenn Rayp, and Marijn Verschelde. 2020. “Firm-Heterogeneous Biased Technological Change: A Nonparametric Approach under Endogeneity.” European Journal of Operational Research 283 (3): 1172–82. https://doi.org/10.1016/j.ejor.2019.11.063.

Dyson, R. G., R. Allen, A. S. Camanho, V. V. Podinovski, C. S. Sarrico, and E. A. Shale. 2001. “Pitfalls and Protocols in DEA.” European Journal of Operational Research 132 (2): 245–59. https://doi.org/10.1016/S0377-2217(00)00149-1.

Eisenhardt, Kathleen M., and Melissa E. Graebner. 2007. “Theory Building from Cases: Opportunities and Challenges.” Academy of Management Journal 50 (1): 25–32. http://amj.aom.org/content/50/1/25.short.

Emrouznejad, Ali, and Guo-liang Yang. 2018. “A Survey and Analysis of the First 40 Years of Scholarly Literature in DEA: 1978–2016.” Socio-Economic Planning Sciences, Recent developments on the use of DEA in the public sector, 61 (March): 4–8. https://doi.org/10.1016/j.seps.2017.01.008.

Färe, Rolf, and Shawna Grosskopf. 2012. Intertemporal Production Frontiers: With Dynamic DEA. Springer Science & Business Media.

Farrell, M. J. 1957. “The Measurement of Productive Efficiency.” Journal of the Royal Statistical Society. Series A (General) 120 (3): 253–90. https://doi.org/10.2307/2343100.

Folkard, Simon, Karen A. Robertson, and Mick B. Spencer. 2007. “A Fatigue/Risk Index to Assess Work Schedules.” Somnologie-Schlafforschung Und Schlafmedizin 11 (3): 177–185.

Frey, Carl Benedikt, and Michael A. Osborne. 2017. “The Future of Employment: How Susceptible Are Jobs to Computerisation?” Technological Forecasting and Social Change 114: 254–280.

Groeneboom, Piet, Geurt Jongbloed, and Jon A. Wellner. 2001. “A Canonical Process for Estimation of Convex Functions: The ‘Invelope’ of Integrated Brownian Motion +t4.” The Annals of Statistics 29 (6): 1620–52. http://www.jstor.org/stable/2699946.

Gstach, Dieter. 1998. “Another Approach to Data Envelopment Analysis in Noisy Environments: DEA+.” Journal of Productivity Analysis 9 (2): 161–76. https://doi.org/10.1023/A:1018312801700.

Hall, Margaret, and Christopher Winsten. 1959. “The Ambiguous Notion of Efficiency.” The Economic Journal 69 (273): 71–86.

Hartigan, J. A., and M. A. Wong. 1979. “Algorithm AS 136: A K-Means Clustering Algorithm.” Journal of the Royal Statistical Society. Series C (Applied Statistics) 28 (1): 100–108. https://doi.org/10.2307/2346830.

Herrera-Restrepo, Oscar, Konstantinos Triantis, William L. Seaver, Joseph C. Paradi, and Haiyan Zhu. 2016. “Bank Branch Operational Performance: A Robust Multivariate and Clustering Approach.” Expert Systems with Applications 50: 107–119. http://www.sciencedirect.com/science/article/pii/S0957417415008271.

Hoff, Ayoe. 2007. “Second Stage DEA: Comparison of Approaches for Modelling the DEA Score.” European Journal of Operational Research 181 (1): 425–35. https://doi.org/10.1016/j.ejor.2006.05.019.

Hubert, Mia, Peter J. Rousseeuw, and Karlien Vanden Branden. 2005. “ROBPCA: A New Approach to Robust Principal Component Analysis.” Technometrics 47 (1): 64–79. https://doi.org/10.1198/004017004000000563.

Johnson, Andrew L., and Timo Kuosmanen. 2011. “One-Stage Estimation of the Effects of Operational Conditions and Practices on Productive Performance: Asymptotically Normal and Efficient, Root-N Consistent StoNEZD Method.” Journal of Productivity Analysis 36 (2): 219–30. http://login.ezproxy.lib.vt.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=eoh&AN=1263102&site=eds-live&scope=site.

Koopmans, Tjalling C. 1951. An Analysis of Production as an Efficient Combination of Activities. Cowles Commission for Research in Economics. New York: John Wiley & Sons.

Krantz, David, Duncan Luce, Patrick Suppes, and Amos Tversky. 1971. “Foundations of Measurement, Vol. I: Additive and Polynomial Representations.”

Kuosmanen, Timo. 2008. “Representation Theorem for Convex Nonparametric Least Squares.” The Econometrics Journal 11 (2): 308–325.

Kuosmanen, Timo, and Andrew L. Johnson. 2009. “Data Envelopment Analysis as Nonparametric Least-Squares Regression.” Operations Research 58 (1): 149–60. https://doi.org/10.1287/opre.1090.0722.

Kuosmanen, Timo, and Mika Kortelainen. 2012. “Stochastic Non-Smooth Envelopment of Data: Semi-Parametric Frontier Estimation Subject to Shape Constraints.” Journal of Productivity Analysis 38 (1): 11–28. https://doi.org/10.1007/s11123-010-0201-3.

Latruffe, Laure, Sophia Davidova, and Kelvin Balcombe. 2008. “Application of a Double Bootstrap to Investigation of Determinants of Technical Efficiency of Farms in Central Europe.” Journal of Productivity Analysis 29 (2): 183–91. https://doi.org/10.1007/s11123-007-0074-2.

Lovell, CA Knox, Lawrence C. Walters, and Lisa L. Wood. 1994. “Stratified Models of Education Production Using Modified DEA and Regression Analysis.” In Data Envelopment Analysis: Theory, Methodology, and Applications, 329–351. Springer.

McDonald, John. 2009. “Using Least Squares and Tobit in Second Stage DEA Efficiency Analyses.” European Journal of Operational Research 197 (2): 792–98. https://doi.org/10.1016/j.ejor.2008.07.039.

Miller, Rupert G. 1974. “The Jackknife--A Review.” Biometrika 61 (1): 1. https://doi.org/10.2307/2334280.

O’Donnell, Christopher J., D. S. Prasada Rao, and George E. Battese. 2008. “Metafrontier Frameworks for the Study of Firm-Level Efficiencies and Technology Ratios.” Empirical Economics 34 (2): 231–55. https://doi.org/10.1007/s00181-007-0119-4.

Olesen, O. B., and N. C. Petersen. 2009. “Target and Technical Efficiency in DEA: Controlling for Environmental Characteristics.” Journal of Productivity Analysis 32 (1): 27–40. https://doi.org/10.1007/s11123-009-0133-y.

Paradi, Joseph C., and H. David Sherman. 2014. “Seeking Greater Practitioner and Managerial Use of DEA for Benchmarking.” Data Envelopment Analysis Journal 1 (1): 29–55. https://www.researchgate.net/profile/Joseph_Paradi/publication/281010454_Seeking_Greater_Practitioner_and_Managerial_Use_of_DEA_for_Benchmarking/links/567eb75d08ae051f9ae655de.pdf.

Peyrache, Antonio, Christiern Rose, and Gabriela Sicilia. 2019. “Variable Selection in Data Envelopment Analysis.” European Journal of Operational Research.

Ray, Subhash C. 1988. “Data Envelopment Analysis, Nondiscretionary Inputs and Efficiency: An Alternative Interpretation.” Socio-Economic Planning Sciences 22 (4): 167–76. https://doi.org/10.1016/0038-0121(88)90003-1.

Roets, Bart, and Johan Christiaens. 2017. “Shift Work, Fatigue and Human Error: An Empirical Analysis of Railway Traffic Control.” Journal of Transportation Safety & Security 0 (ja): 1–18. https://doi.org/10.1080/19439962.2017.1376022.

Ruggiero, John. 1996. “On the Measurement of Technical Efficiency in the Public Sector.” European Journal of Operational Research 90 (3): 553–65. https://doi.org/10.1016/0377-2217(94)00346-7.

Seaver, Bill L., and Konstantinos P. Triantis. 1992. “A Fuzzy Clustering Approach Used in Evaluating Technical Efficiency Measures in Manufacturing.” Journal of Productivity Analysis 3 (4): 337–363. http://www.springerlink.com/index/x05527653768225m.pdf.

Simar, Léopold, and Paul W. Wilson. 2007. “Estimation and Inference in Two-Stage, Semi-Parametric Models of Production Processes.” Journal of Econometrics 136 (1): 31–64. https://doi.org/10.1016/j.jeconom.2005.07.009.

———. 2011. “Two-Stage DEA: Caveat Emptor.” Journal of Productivity Analysis 36 (2): 205. https://doi.org/10.1007/s11123-011-0230-6.

———. 2015. “Statistical Approaches for Non-Parametric Frontier Models: A Guided Tour.” International Statistical Review 83 (1): 77–110.

Stanton, Kenneth R. 2002. “Trends in Relationship Lending and Factors Affecting Relationship Lending Efficiency.” Journal of Banking & Finance 26 (1): 127–152.

Topcu, Taylan G., Konstantinos Triantis, and Bart Roets. 2019. “Estimation of the Workload Boundary in Socio-Technical Infrastructure Management Systems: The Case of Belgian Railroads.” European Journal of Operational Research 278 (1): 314–29. https://doi.org/10.1016/j.ejor.2019.04.009.

Triantis, K. 2015. “Engineering Design and Efficiency Measurement: Issues and Future Research Opportunities.” Data Envelopment Analysis Journal 1 (2): 81–112. http://econpapers.repec.org/RePEc:now:jnldea:103.00000008.

Triantis, Konstantinos, Devang Sarayia, and Bill Seaver. 2010. “Using Multivariate Methods to Incorporate Environmental Variables for Local and Global Efficiency Performance Analysis.” INFOR: Information Systems and Operational Research 48 (1): 39–52. https://doi.org/10.3138/infor.48.1.039.

Turner, Hugh, Robert Windle, and Martin Dresner. 2004. “North American Containerport Productivity: 1984–1997.” Transportation Research Part E: Logistics and Transportation Review 40 (4): 339–356.

Verschelde, Marijn, and Nicky Rogge. 2012. “An Environment-Adjusted Evaluation of Citizen Satisfaction with Local Police Effectiveness: Evidence from a Conditional Data Envelopment Analysis Approach.” European Journal of Operational Research 223 (1): 214–25. https://doi.org/10.1016/j.ejor.2012.05.044.

Vidoli, Francesco, and Giancarlo Ferrara. 2015. “Analyzing Italian Citrus Sector by Semi-Nonparametric Frontier Efficiency Models.” Empirical Economics 49 (2): 641–58. https://doi.org/10.1007/s00181-014-0879-6.

Weick, Karl E. 1995. “What Theory Is Not, Theorizing Is.” Administrative Science Quarterly, 385–390. http://www.jstor.org/stable/2393789.

Wong, M. Anthony, and Tom Lane. 1983. “A Kth Nearest Neighbour Clustering Procedure.” Journal of the Royal Statistical Society. Series B (Methodological) 45 (3): 362–68. http://www.jstor.org/stable/2345405.

Chapter 5. Conclusions

The reliance on autonomous technologies will only continue to increase. While increased autonomy has the potential for significantly improving our lives, many autonomous systems will rely on people for supervision, at least for a foreseeable future. The complicated nature of the human mind and its susceptibility to a wide-range of contextual influences require us to further investigate how these social and technical elements can be integrated as a cohesive whole. Therefore, this dissertation focused on the interdependencies between human decision-makers and their autonomous counterparts that operate within sociotechnical enterprises.

When I started my doctoral studies, I was convinced that one needs to investigate operational systems to (i) learn about their complex behavior and (ii) be able verify the obtained results. I believed that this would allow one to eliminate the need for crude simplifications and help to understand how extreme cases come to occurrence. Thus, this dissertation provided an empirical mixed-methods approach that brought together theory with practice to further our understanding of sociotechnical systems. More specifically, this dissertation focused on three aspects of STSs, safe management, collaboration, and efficiency measurement. While the results of this dissertation are far from being conclusive, they are useful because they are verified by the operations of a real-world STS.

5.1 Chapter 2 - Safe Management

Chapter 2 was primarily motivated by some of the recent airline tragedies that resulted in hundreds of fatalities (Salmon, Walker, and Stanton 2016; Tjahjono 2018; Johnston and Harris 2019; The Aircraft Accident Investigation Bureau of Ethiopia 2019). Like many others, these accidents were caused by a combination of engineered system failures and the inability of human controllers to prevent catastrophic consequences. I considered these accidents an indicator of how increased autonomy could introduce unanticipated failure modes that could be unique and difficult to prevent. Rasmussen previously theorized that these hard to interpret failure modes could be considered in three distinct modes (Rasmussen 1997; Cook and Rasmussen 2005), one of which is the workload boundary that represent the amount of work allocated on safety-critical decision-makers. The first essay quantified this workload boundary that remained a qualitative construct for over two

decades, through a microeconomic approach that assumed it could be estimated through a Pareto-Koopmans Frontier (Koopmans 1951; Farrell 1957).

I identified these contextual influences through an interdisciplinary literature review and formulated a multivariate clustering approach (Triantis, Sarayia, and Seaver 2010) to quantify the workload boundary on a pilot INFRABEL traffic control center. The employed analytical technique allowed me to quantify the aggregate impact of contextual factors, which was observed to be significant with sometimes up to 60% of the observable workload. The approach was verified by domain experts and was implemented on-site. To the best of my knowledge, this is the first quantification of Rasmussen's workload boundary and it provides a novel, holistic, and systems oriented mechanism for practitioners to manage their STSs safely.

5.2 Chapter 3 - Collaboration

Chapter 3 was primarily concerned with how teams of humans and their autonomous partners share work, given their subjective preferences and contextual operational conditions. I extended the workload measurement approach that was proposed in the second chapter, to understand how the amount of work carried out by each agent varied. Since the dataset included a high fraction of influential observations, I utilized the prediction power of machine learning algorithms, and proposed a novel integration with efficiency measurement techniques. Machine learning techniques were previously never used to explain contextual influences within an efficiency measurement framework. The demonstrated approach reveals the preferences of human decision-makers, without interfering with their daily operational behavior.

The results of this study supported the literature in terms of documenting that to a great extent, autonomous systems are preferred to handle low complexity and density tasks. I observed intensifying operational conditions increase the reliance on collaboration, which was indicated by the rise in the workloads of both human and autonomous agents. I observed great variability in how people chose to collaborate with their autonomous counterparts, which led me to believe that (i) the problem is much more complex and (ii) there could be psychological and behavioral root causes, which one may not be able to capture with the kind of abstraction techniques that were used in this dissertation.

5.3 Chapter 3 – Efficiency Measurement

The third essay explored the complementary and contrasting roles of analytical efficiency measurement approaches that deal with the influence of contextual factors and their sensitivity to sample size. More specifically, I implemented the popular two-stage approach of Simar and Wilson and compared its relative insights with respect to the multivariate clustering approach. I observed that, strong restrictive assumptions, such as the separability condition, could be restrictive and may not apply to complex STSs. Results of this essay raised caution to the vast number of studies that disregarded these considerations and supported previous concerns (Olesen and Petersen 2009; Bădin, Daraio, and Simar 2010; Daraio, Simar, and Wilson 2018; Banker, Natarajan, and Zhang 2019). I organized the results in a structured taxonomy based on their fundamental assumptions, limitations, mathematical structure, sensitivity to sample size, and their practical usefulness.

5.4 Future Work

This dissertation left more questions unanswered than the ones it addressed. Below I discuss some of the directions this research could take in the future:

About Rasmussen's Safe Operation Envelope for STSs: While the quantification of Rasmussen's workload boundary through a Pareto-Koopmans frontier made sense in this specific application, which is encouraging, it remains to be seen if it is generalizable to other application areas. It would be interesting to see if it could be extended to a similar system, e.g. air traffic control, and still make sense for practitioners and domain experts. If it does indeed make sense in a different context, perhaps it could serve as a rigorous management approach for STSs in general.

Regardless of generalizability, this naïve quantification allowed me to document the underload-overload cycles that occur throughout the day and helped to measure how drastically these cycles could differ. Underload cycles lead to distraction and loss of attention, which in return contributes to STS accidents, yet this dissertation solely focused on the overload. This raises question on some of the other assumptions made in this dissertation. For example, I assumed that a pure technical efficiency on the workload

boundary, represented by an efficiency score of one (which means overload), indicates a system level risk. This assumption needs further verification. Perhaps, one could focus on the occurrences of near-miss events (Dillon et al. 2016) and their relationship to workload to test this assumption. More importantly, one would expect the relationship of the workload boundary with the other two boundaries to play an important role on the manifestation of near-miss events. Another assumption was that all observed decisions were of equal value. Differentiating between good and bad decisions, would necessitate an in-depth investigation, therefore could benefit from a different research strategy. Which brings us to the second future research direction.

Human – Autonomous System Collaboration: While I explored collaboration in STSs given contextual influences, the scope of available data didn't allow to explore the specific interactions that lead a decision-maker to delegate authority or retain authority from their autonomous counterparts. This led me to formulate a mutually exclusive measurement model, while clearly these two agents (human and autonomous) are working dependently. Unfortunately, none of our datasets included social and behavioral considerations such as beliefs, trust, previous experiences, educational level, attitude towards automation etc. What are the factors that lead a person to trust an autonomous system? Is trust a process? Do people use autonomous systems because they need them? How can we design trustworthy autonomous systems? These research questions could be addressed in the future.

Moreover, this dissertation did not focus on cognitive and/or behavioral patterns that are inherent to a specific team of people who work in the same location. Similarly, there could be hourly, daily, or organizational tendencies. One particular simplification I made (by using a DEA model) was to falsely assume that the investigated decision-makers do not interact with one another and influence their decisions. Obviously, this is not true, controllers frequently interact, and further research could highlight important knowledge gaps. This brings us to the next future research direction.

Organizational Impact of Increased Autonomy: This dissertation was solely focused on the relationship between a human decision maker and its autonomous counterpart, and the relationship with the rest of the organization was purposefully left out of scope. One

particular simplification I had to make was to disregard how the collaborating human-autonomous system team was managed by a higher-level supervisor. Conflict resolution and coordination within a sociotechnical enterprise would be an interesting venue to explore.

Additionally, this dissertation did not investigate the impact of automation on the performance of an enterprise or an organization. Are increasing levels of automation are really helping organizations to reach their long-term goals? Does increasing levels of automation help to minimize unintended consequences and improve overall performance? If not, why and how it can be enhanced? These questions bring me to the another research direction this dissertation could possibly take.

Systems Engineering Sociotechnical Organizations: This dissertation investigated sociotechnical systems without opening up the black-box of automation. Without delving into the details of how autonomous systems make their decisions, I firmly believe it might not be possible to understand the complex nature of STSs. If we precisely knew which sources of information a human decision-maker seeks given a situation, would it allow us to improve the design of the engineered elements? From this perspective, understanding the contextual influences on human decisions is useful yet insufficient. There is a need to capture the sensory interactions, e.g. eye tracking, that take place through the interface between the human and the autonomous unit. Would re-designing the interface based on interactions and their contextual variance improve performance? If so by how much?

Assuming that increasing system autonomy will lead to increased centralization (less people, more activities to oversee & control), there will be a need to design the architecture of sociotechnical systems where people are reconsidered as subsystems. This could potentially allow to strategically design and utilize the strengths of both social and technical spheres. One could potentially explore, system architecture techniques that allow to investigate the tradeoffs between people and autonomous systems. Is there an automation sweet-spot? How can one re-arrange existing relationships within an organization, or incorporate these relationships into the design of an engineered artifact to enhance performance?

References

Acemoglu, Daron, and Pascual Restrepo. 2017. “Robots and Jobs: Evidence from US Labor Markets.” NBER Working Paper No. W23285.

Alge, Bradley J., and S. Duane Hansen. 2014. Workplace Monitoring and Surveillance Research since 1984: A Review and Agenda. Routledge, New York.

Ali, Jehad, Rehanullah Khan, Nasir Ahmad, and Imran Maqsood. 2012. “Random Forests and Decision Trees.” International Journal of Computer Science Issues (IJCSI); Mahebourg 9 (5): 272–78.

Alpaydin, Ethem. 2009. Introduction to Machine Learning. MIT press.

Altmann, André, Laura Tolosi, Oliver Sander, and Thomas Lengauer. 2010. “Data and Text Mining Permutation Importance: A Corrected Feature Importance Measure.” Bioinformatics 26 (10): 1340–47. https://doi.org/10.1093/bioinformatics/btq134.

Aragon, Y., A. Daouia, and C. Thomas-Agnan. 2005. “Nonparametric Frontier Estimation: A Conditional Quantile-Based Approach.” Econometric Theory 21 (2): 358–89. https://doi.org/10.1017/S0266466605050206.

Arnaldo, Rosa M., V. Fernando Gómez Comendador, Rocio Barragan, and Luis Pérez. 2014. “European Air Navigation Service Providers’ Efficiency Evaluation through Data Envelopment Analysis (DEA).” In 29th Congress of the International Council of the Aeronautical Sciences.

Athanassopoulos, Antreas D., and Stephen P. Curram. 1996. “A Comparison of Data Envelopment Analysis and Artificial Neural Networks as Tools for Assessing the Efficiency of Decision Making Units.” The Journal of the Operational Research Society 47 (8): 1000–1016. https://doi.org/10.2307/3010408.

Athanassopoulos, Antreas, and Konstantions Triantis. 1998. “Assessing Aggregate Cost Efficiency and the Related Policy Implications for Greek Local Municipalities - ProQuest.” INFOR: Information Systems and Operational Research 36 (3): 66–83.

Azadeh, A., Z. Gaeini, S. Motevali Haghighi, and B. Nasirian. 2016. “A Unique Adaptive Neuro Fuzzy Inference System for Optimum Decision Making Process in a Natural Gas Transmission Unit.” Journal of Natural Gas Science and Engineering 34 (August): 472–85. https://doi.org/10.1016/j.jngse.2016.06.053.

Azadeh, A., M. Hasannia Kolaee, and M. Sheikhalishahi. 2016. “An Integrated Approach for Configuration Optimization in a CBM System by Considering Fatigue Effects.” The International Journal of Advanced Manufacturing Technology 86 (5–8): 1881–93. https://doi.org/10.1007/s00170-015-8204-x.

Azadeh, A., H. Tohidi, M. Zarrin, S. Pashapour, and M. Moghaddam. 2016. “An Integrated Algorithm for Performance Optimization of Neurosurgical ICUs.” Expert Systems with Applications 43 (Supplement C): 142–53. https://doi.org/10.1016/j.eswa.2015.08.042.

Azadeh, Ali, Morteza Saberi, Reza Tavakkoli Moghaddam, and Leili Javanmardi. 2011. “An Integrated Data Envelopment Analysis–Artificial Neural Network–Rough Set Algorithm for Assessment of Personnel Efficiency.” Expert Systems with Applications 38 (3): 1364–73. https://doi.org/10.1016/j.eswa.2010.07.033.

Babajide Mustapha, Ismail, and Faisal Saeed. 2016. “Bioactive Molecule Prediction Using Extreme Gradient Boosting.” Molecules 21 (8): 983. https://doi.org/10.3390/molecules21080983.

Bădin, Luiza, Cinzia Daraio, and Léopold Simar. 2010. “Optimal Bandwidth Selection for Conditional Efficiency Measures: A Data-Driven Approach.” European Journal of Operational Research 201 (2): 633–40. https://doi.org/10.1016/j.ejor.2009.03.038.

Bădin, Luiza, Cinzia Daraio, and Léopold Simar. 2014. “Explaining Inefficiency in Nonparametric Production Models: The State of the Art.” Annals of Operations Research 214 (1): 5–30. https://doi.org/10.1007/s10479-012-1173-7.

Badin, Luiza, Leopold Simar, and Cinzia Daraio. 2012. “How to Measure the Impact of Environmental Factors in a Nonparametric Production Model.” European Journal of Operational Research 223 (3): 818–33.

Banker, Rajiv D., Abraham Charnes, and William Wager Cooper. 1984. “Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis.” Management Science 30 (9): 1078–1092.

Banker, Rajiv D., and Richard C. Morey. 1986. “The Use of Categorical Variables in Data Envelopment Analysis.” Management Science 32 (12): 1613–1627.

Banker, Rajiv D., and Ram Natarajan. 2008. “Evaluating Contextual Variables Affecting Productivity Using Data Envelopment Analysis.” Operations Research 56 (1): 48–58. http://www.jstor.org.ezproxy.lib.vt.edu/stable/25147166.

Banker, Rajiv, Ram Natarajan, and Daqun Zhang. 2019. “Two-Stage Estimation of the Impact of Contextual Variables in Stochastic Frontier Production Function Models Using Data Envelopment Analysis: Second Stage OLS versus Bootstrap Approaches.” European Journal of Operational Research, Advances in Data Envelopment Analysis, 278 (2): 368–84. https://doi.org/10.1016/j.ejor.2018.10.050.

Barling, Julian, Catherine Loughlin, and E. Kevin Kelloway. 2002. “Development and Test of a Model Linking Safety-Specific Transformational Leadership and Occupational Safety.” Journal of Applied Psychology 87 (3): 488.

Barros, Carlos Pestana, Milton Nektarios, and A. Assaf. 2010. “Efficiency in the Greek Insurance Industry.” European Journal of Operational Research 205 (2): 431–36. https://doi.org/10.1016/j.ejor.2010.01.011.

Battese, George E., D. S. Prasada Rao, and Christopher J. O’Donnell. 2004. “A Metafrontier Production Function for Estimation of Technical Efficiencies and Technology Gaps for Firms Operating Under Different Technologies.” Journal of Productivity Analysis 21 (1): 91–103. https://doi.org/10.1023/B:PROD.0000012454.06094.29.

Baxter, Gordon, and Ian Sommerville. 2011. “Socio-Technical Systems: From Design Methods to Systems Engineering.” Interacting with Computers 23 (1): 4–17.

Beehr, Terry A. 2014. Psychological Stress in the Workplace (Psychology Revivals). Routledge.

Bergstra, James, and Yoshua Bengio. 2012. “Random Search for Hyper-Parameter Optimization.” Journal of Machine Learning Research 13 (Feb): 281–305.

Besstremyannaya, Galina. 2011. “Managerial Performance and Cost Efficiency of Japanese Local Public Hospitals: A Latent Class Stochastic Frontier Model.” Health Economics 20 (S1): 19–34.

Blank, Jos L. T., and Vivian G. Valdmanis. 2010. “Environmental Factors and Productivity on Dutch Hospitals: A Semi-Parametric Approach.” Health Care Management Science 13 (1): 27–34. https://doi.org/10.1007/s10729-009-9104-0.

Bogetoft, Peter, and Lars Otto. 2010. Benchmarking with DEA, SFA, and R. International Series in Operations Research & Management Science. Springer.

Borchani, Hanen, Gherardo Varando, Concha Bielza, and Pedro Larrañaga. 2015. “A Survey on Multi-Output Regression.” Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 5 (5): 216–233.

Borja, Alexandra Medina, and Konstantinos Triantis. 2007. “A Conceptual Framework to Evaluate Performance of Non-Profit Social Service Organisations.” International Journal of Technology Management 37 (1/2): 147. https://doi.org/10.1504/IJTM.2007.011808.

Boser, Bernhard E, Isabelle M Guyon, and Vladimir N Vapnik. 1992. “A Training Algorithm for Optimal Margin Classiiers.” Proceedings of the Fifth Annual Workshop on Computational Learning Theory ACM.

Breiman, Leo. 2001. “Random Forests.” Machine Learning 45 (1): 5–32.

Breiman, Leo. 2017. Classification and Regression Trees. Routledge. https://doi.org/10.1201/9781315139470.

Broniatowski, David A., and Conrad Tucker. 2017. “Assessing Causal Claims about Complex Engineered Systems with Quantitative Data: Internal, External, and Construct Validity.” Systems Engineering 20 (6): 483–96. https://doi.org/10.1002/sys.21414.

Buede, Dennis M., and William D. Miller. 2016. The Engineering Design of Systems: Models and Methods. John Wiley & Sons.

Carayon, Pascale, Peter Hancock, Nancy Leveson, Ian Noy, Laerte Sznelwar, and Geert van Hootegem. 2015. “Advancing a Sociotechnical Systems Approach to Workplace Safety – Developing the Conceptual Framework.” Ergonomics 58 (4): 548–64. https://doi.org/10.1080/00140139.2015.1015623.

Cazals, Catherine, Jean-Pierre Florens, and Léopold Simar. 2002. “Nonparametric Frontier Estimation: A Robust Approach.” Journal of Econometrics 106 (1): 1–25. https://doi.org/10.1016/S0304-4076(01)00080-X.

Chambers, Christopher P., and Federico Echenique. 2016. Revealed Preference Theory. Vol. 56. Cambridge University Press.

Charnes, Abraham, William W. Cooper, Boaz Golany, Larry Seiford, and J. Stutz. 1985. “Foundations of Data Envelopment Analysis for Pareto-Koopmans Efficient Empirical Production Functions.” Journal of Econometrics 30 (1–2): 91–107.

Charnes, Abraham, William W. Cooper, and Edwardo Rhodes. 1978. “Measuring the Efficiency of Decision Making Units.” European Journal of Operational Research 2(6):429–444.

Chen, Jessie Y. C., and Michael J. Barnes. 2014. “Human–Agent Teaming for Multirobot Control: A Review of Human Factors Issues.” IEEE Transactions on Human-Machine Systems 44 (1): 13–29. https://doi.org/10.1109/THMS.2013.2293535.

Chen, Tianqi, and Carlos Guestrin. 2016. “XGBoost: A Scalable Tree Boosting System.” In Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. KDD ’16. New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785.

Chilingerian, Jon A., and H. David Sherman. 2004. “Health Care Applications.” In Handbook on Data Envelopment Analysis, 481–537. Springer.

Chollet, Francois. 2015. Keras. https://keras.io.

Cook, R., and J. Rasmussen. 2005. “‘Going Solid’: A Model of System Dynamics and Consequences for Patient Safety.” BMJ Quality & Safety 14 (2): 130–34. https://doi.org/10.1136/qshc.2003.009530.

Corley, Kevin G., and Dennis A. Gioia. 2011. “Building Theory about Theory Building: What Constitutes a Theoretical Contribution?” Academy of Management Review 36 (1): 12–32. http://amr.aom.org/content/36/1/12.1.short.

Cortes, Corinna, and Vladimir Vapnik. 1995. “Support-Vector Networks.” Machine Learning 20 (3): 273–97.

Costa, Álvaro, and Raphael N. Markellos. 1997. “Evaluating Public Transport Efficiency with Neural Network Models.” Transportation Research Part C: Emerging Technologies 5 (5): 301–12. https://doi.org/10.1016/S0968-090X(97)00017-X.

Ćujić, Mara, Milica Jovanović, Gordana Savić, and Maja Levi Jakšić. 2015. “Measuring the Efficiency of Air Navigation Services System by Using DEA Method.” International Journal for Traffic and Transport Engineering 5 (1).

Cullen, Jennifer C., and Leslie B. Hammer. 2007. “Developing and Testing a Theoretical Model Linking Work-Family Conflict to Employee Safety.” Journal of Occupational Health Psychology 12 (3): 266. http://psycnet.apa.org/journals/ocp/12/3/266/.

Dai, Xiaofeng, and Timo Kuosmanen. 2014. “Best-Practice Benchmarking Using Clustering Methods: Application to Energy Regulation.” Omega 42 (1): 179–88. https://doi.org/10.1016/j.omega.2013.05.007.

Daraio, Cinzia, and Léopold Simar. 2005. “Introducing Environmental Variables in Nonparametric Frontier Models: A Probabilistic Approach.” Journal of Productivity Analysis 24 (1): 93–121. https://doi.org/10.1007/s11123-005-3042-8.

Daraio, Cinzia, Léopold Simar, and Paul W. Wilson. 2010. “Testing Whether Two-Stage Estimation Is Meaningful in Non-Parametric Models of Production.” ISBA Discussion Paper.

Daraio, Cinzia, Léopold Simar, and Paul W. Wilson. 2018. “Central Limit Theorems for Conditional Efficiency Measures and Tests of the ‘Separability’ Condition in Non-Parametric, Two-Stage Models of Production.” The Econometrics Journal 21 (2): 170–91. https://doi.org/10.1111/ectj.12103.

Dawson, Drew, and Kirsty McCulloch. 2005. “Managing Fatigue: It’s about Sleep.” Sleep Medicine Reviews 9 (5): 365–380.

De Bruijn, Hans, and Paulien M. Herder. 2009. “System and Actor Perspectives on Sociotechnical Systems.” IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans 39 (5): 981–992.

Debreu, Gerard. 1951. “The Coefficient of Resource Utilization.” Econometrica 19 (3): 273–92. https://doi.org/10.2307/1906814.

Dekker, Sidney. 2016. Drift into Failure: From Hunting Broken Components to Understanding Complex Systems. CRC Press.

Department of Defense. 2012. “Department of Defense Standard Practice System Safety.” MIL-STD-882E.

Deprins, Dominique, Léopold Simar, and Henry Tulkens. 2006. “Measuring Labor-Efficiency in Post Offices.” In Public Goods, Environmental Externalities and Fiscal Competition, edited by Parkash Chander, Jacques Drèze, C. Knox Lovell, and Jack Mintz, 285–309. Boston, MA: Springer US. https://doi.org/10.1007/978-0-387-25534-7_16.

Dewitte, Ruben, Michel Dumont, Bruno Merlevede, Glenn Rayp, and Marijn Verschelde. 2020. “Firm-Heterogeneous Biased Technological Change: A Nonparametric Approach under Endogeneity.” European Journal of Operational Research 283 (3): 1172–82. https://doi.org/10.1016/j.ejor.2019.11.063.

Dillon, Robin L., Catherine H. Tinsley, Peter M. Madsen, and Edward W. Rogers. 2016. “Organizational Correctives for Improving Recognition of Near-Miss Events.” Journal of Management 42 (3): 671–97. https://doi.org/10.1177/0149206313498905.

Dismukes, R. Key. 2012. “Prospective Memory in Workplace and Everyday Situations.” Current Directions in Psychological Science 21 (4): 215–220.

Dorrian, Jillian, Stuart D. Baulk, and Drew Dawson. 2011. “Work Hours, Workload, Sleep and Fatigue in Australian Rail Industry Employees.” Applied Ergonomics, Special Section: Ergonomics, health and working time organization, 42 (2): 202–9. https://doi.org/10.1016/j.apergo.2010.06.009.

Dyson, R. G., R. Allen, A. S. Camanho, V. V. Podinovski, C. S. Sarrico, and E. A. Shale. 2001. “Pitfalls and Protocols in DEA.” European Journal of Operational Research 132 (2): 245–59. https://doi.org/10.1016/S0377-2217(00)00149-1.

Eisenhardt, Kathleen M., and Melissa E. Graebner. 2007. “Theory Building from Cases: Opportunities and Challenges.” Academy of Management Journal 50 (1): 25–32. http://amj.aom.org/content/50/1/25.short.

El-Mahgary, Sami, and Risto Lahdelma. 1995. “Data Envelopment Analysis: Visualizing the Results.” European Journal of Operational Research 83 (3): 700–710. https://doi.org/10.1016/0377-2217(94)00303-T.

Emrouznejad, Ali, and Guo-liang Yang. 2018. “A Survey and Analysis of the First 40 Years of Scholarly Literature in DEA: 1978–2016.” Socio-Economic Planning Sciences, Recent developments on the use of DEA in the public sector, 61 (March): 4–8. https://doi.org/10.1016/j.seps.2017.01.008.

Färe, Rolf, and Shawna Grosskopf. 2012. Intertemporal Production Frontiers: With Dynamic DEA. Springer Science & Business Media.

Farrell, M. J. 1957. “The Measurement of Productive Efficiency.” Journal of the Royal Statistical Society. Series A (General) 120 (3): 253–90. https://doi.org/10.2307/2343100.

Farrington-Darby, T., John R. Wilson, B. J. Norris, and Theresa Clarke. 2006. “A Naturalistic Study of Railway Controllers.” Ergonomics 49 (12–13): 1370–94. https://doi.org/10.1080/00140130600613000.

Ferguson, Sally A., Nicole Lamond, Katie Kandelaars, Sarah M. Jay, and Drew Dawson. 2008. “The Impact of Short, Irregular Sleep Opportunities at Sea on the Alertness of Marine Pilots Working Extended Hours.” Chronobiology International 25 (2–3): 399–411.

Folkard, Simon, Karen A. Robertson, and Mick B. Spencer. 2007. “A Fatigue/Risk Index to Assess Work Schedules.” Somnologie-Schlafforschung Und Schlafmedizin 11 (3): 177–185.

Franssen, Maarten. 2005. “Arrow’s Theorem, Multi-Criteria Decision Problems and Multi-Attribute Preferences in Engineering Design.” Research in Engineering Design 16 (1–2): 42–56. https://doi.org/10.1007/s00163-004-0057-5.

Frey, Carl Benedikt, and Michael A. Osborne. 2017. “The Future of Employment: How Susceptible Are Jobs to Computerisation?” Technological Forecasting and Social Change 114: 254–280.

Friedman, Jerome H. 2001. “Greedy Function Approximation: A Gradient Boosting Machine.” The Annals of Statistics 29 (5): 1189–1232. https://doi.org/10.1214/aos/1013203451.

Friedman, Jerome H. 2002. “Stochastic Gradient Boosting.” Computational Statistics & Data Analysis 38 (4): 367–78. https://doi.org/10.1016/S0167-9473(01)00065-2.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2000. “Additive Logistic Regression: A Statistical View of Boosting (With Discussion and a Rejoinder by the Authors).” The Annals of Statistics 28 (2): 337–407. https://doi.org/10.1214/aos/1016218223.

Gardner, M W, and S R Dorling. 1998. “ARTIFICIAL NEURAL NETWORKS (THE MULTILAYER PERCEPTRON)-A REVIEW OF APPLICATIONS IN THE ATMOSPHERIC SCIENCES.” Atmospheric Environment. Vol. 32.

Golany, B, and Y Roll. 1989. “An Application Procedure for DEA.” Omega 17 (3): 237–50. https://doi.org/10.1016/0305-0483(89)90029-7.

Gorman, Jamie C., Nancy J. Cooke, Eduardo Salas, and Barry Strauch. 2010. “Can Cultural Differences Lead to Accidents? Team Cultural Differences and Sociotechnical System Operations.” Human Factors 52 (2): 246–63. https://doi.org/10.1177/0018720810362238.

Groeneboom, Piet, Geurt Jongbloed, and Jon A. Wellner. 2001. “A Canonical Process for Estimation of Convex Functions: The ‘Invelope’ of Integrated Brownian Motion +t4.” The Annals of Statistics 29 (6): 1620–52. http://www.jstor.org/stable/2699946.

Grundgeiger, Tobias, Penelope M. Sanderson, and R. Key Dismukes. 2015. “Prospective Memory in Complex Sociotechnical Systems.” Zeitschrift Für Psychologie.

Gstach, Dieter. 1998. “Another Approach to Data Envelopment Analysis in Noisy Environments: DEA+.” Journal of Productivity Analysis 9 (2): 161–76. https://doi.org/10.1023/A:1018312801700.

Guyon, Isabelle, B Boser, and Vladimir Vapnik. 1993. “Automatic Capacity Tuning of Very Large VC-Dimension Classifiers.” In Advances in Neural Information Processing Systems, 147–55.

Hall, Margaret, and Christopher Winsten. 1959. “The Ambiguous Notion of Efficiency.” The Economic Journal 69 (273): 71–86.

Hampel, Frank R., Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. 2011. Robust Statistics: The Approach Based on Influence Functions. John Wiley & Sons.

Hartigan, J. A., and M. A. Wong. 1979. “Algorithm AS 136: A K-Means Clustering Algorithm.” Journal of the Royal Statistical Society. Series C (Applied Statistics) 28 (1): 100–108. https://doi.org/10.2307/2346830.

Hazelrigg, George. 1998. “A Framework for Decision-Based Engineering Design.” Journal of Mechanical Design 120 (4): 653–58. https://doi.org/10.1115/1.2829328.

Hendrick, Hal W. 1995. “Humanizing Re-Engineering for True Organizational Effectiveness: A Macroergonomic Approach.” Proceedings of the Human Factors and Ergonomics Society Annual Meeting 39 (12): 761–65. https://doi.org/10.1177/154193129503901202.

Herrera-Restrepo, O., and K. Triantis. 2018. “Efficiency-Driven Enterprise Design: A Synthesis of Studies.” IEEE Transactions on Engineering Management 65 (3): 363–78. https://doi.org/10.1109/TEM.2018.2795563.

Herrera-Restrepo, Oscar, and Konstantinos Triantis. 2019. “Enterprise Design through Complex Adaptive Systems and Efficiency Measurement.” European Journal of Operational Research, Advances in Data Envelopment Analysis, 278 (2): 481–97. https://doi.org/10.1016/j.ejor.2018.12.002.

Herrera-Restrepo, Oscar, Konstantinos Triantis, William L. Seaver, Joseph C. Paradi, and Haiyan Zhu. 2016a. “Bank Branch Operational Performance: A Robust Multivariate and Clustering Approach.” Expert Systems with Applications 50: 107–119.

Heydari, Babak, Zoe Szajnfarber, Jitesh Panchal, Michel-Alexandre Cardin, Katja Holtta-Otto, Gül E. Kremer, and Wei Chen. 2019. “Special Issue: Analysis and Design of Sociotechnical Systems.” Journal of Mechanical Design 141 (11). https://doi.org/10.1115/1.4029150.

Hodgson, Allan, Carys E. Siemieniuch, and Ella-Mae Hubbard. 2013. “Culture and the Safety of Complex Automated Sociotechnical Systems.” IEEE Transactions on Human-Machine Systems 43 (6): 608–19. https://doi.org/10.1109/THMS.2013.2285048.

Hoff, Ayoe. 2007. “Second Stage DEA: Comparison of Approaches for Modelling the DEA Score.” European Journal of Operational Research 181 (1): 425–35. https://doi.org/10.1016/j.ejor.2006.05.019.

Hofmann, David A., and Frederick P. Morgeson. 1999. “Safety-Related Behavior as a Social Exchange: The Role of Perceived Organizational Support and Leader–Member Exchange.” Journal of Applied Psychology 84 (2): 286. http://psycnet.apa.org/journals/apl/84/2/286/.

Hornik, Kurt, Maxwell Stinchcombe, and Halbert White. 1989. “Multilayer Feedforward Networks Are Universal Approximators.” Neural Networks 2 (5): 359–66. https://doi.org/10.1016/0893-6080(89)90020-8.

Hubert, Mia, Peter J. Rousseeuw, and Karlien Vanden Branden. 2005. “ROBPCA: A New Approach to Robust Principal Component Analysis.” Technometrics 47 (1): 64–79. https://doi.org/10.1198/004017004000000563.

Jain, Anil K., Jianchang Mao, and K. M. Mohiuddin. 1996. “Artificial Neural Networks: A Tutorial.” Computer, no. 3: 31–44.

Jain, Sanjay, Konstantinos P. Triantis, and Shiyong Liu. 2011. “Manufacturing Performance Measurement and Target Setting: A Data Envelopment Analysis Approach.” European Journal of Operational Research 214 (3): 616–26. https://doi.org/10.1016/j.ejor.2011.05.028.

Johnson, Andrew L., and Timo Kuosmanen. 2011. “One-Stage Estimation of the Effects of Operational Conditions and Practices on Productive Performance: Asymptotically Normal and Efficient, Root-N Consistent StoNEZD Method.” Journal of Productivity Analysis 36 (2): 219–30.

Johnson, Andrew L., and Timo Kuosmanen. 2012. “One-Stage and Two-Stage DEA Estimation of the Effects of Contextual Variables.” European Journal of Operational Research 220 (2): 559–70. https://doi.org/10.1016/j.ejor.2012.01.023.

Johnston, Phillip, and Rozi Harris. 2019. “The Boeing 737 MAX Saga: Lessons for Software Organizations.” Software Quality Professional; Milwaukee 21 (3): 4–12

Karlaftis, M. G., and E. I. Vlahogianni. 2011. “Statistical Methods versus Neural Networks in Transportation Research: Differences, Similarities and Some Insights.” Transportation Research Part C: Emerging Technologies 19 (3): 387–99. https://doi.org/10.1016/j.trc.2010.10.004.

Keeney, Ralph L., and Howard Raiffa. 1993. Decisions with Multiple Objectives: Preferences and Value Trade-Offs. Cambridge University Press.

Kerns, Kimberly A. 2000. “The CyberCruiser: An Investigation of Development of Prospective Memory in Children.” Journal of the International Neuropsychological Society 6 (1): 62–70.

Khezrimotlagh, Dariush, Joe Zhu, Wade D. Cook, and Mehdi Toloo. 2019. “Data Envelopment Analysis and Big Data.” European Journal of Operational Research 274 (3): 1047–54. https://doi.org/10.1016/j.ejor.2018.10.044.

Kleiner, Brian M., Lawrence J. Hettinger, David M. DeJoy, Yuang-Hsiang Huang, and Peter E. D. Love. 2015a. “Sociotechnical Attributes of Safe and Unsafe Work Systems.” Ergonomics 58 (4): 635–49. https://doi.org/10.1080/00140139.2015.1009175.

Kohavi, Ron. 1995. “A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection.” In Ijcai, 14:1137–1145. Montreal, Canada.

Koopmans, Tjalling C. 1951. An Analysis of Production as an Efficient Combination of Activities. Cowles Commission for Research in Economics. New York: John Wiley & Sons.

Kopardekar, Parimal, and Sherri Magyarits. 2002. “Dynamic Density: Measuring and Predicting Sector Complexity [ATC].” In Digital Avionics Systems Conference, 2002. Proceedings. The 21st, 1:2C4–2C4. IEEE.

Kotsiantis, Sotiris B., I. Zaharakis, and P. Pintelas. 2007. “Supervised Machine Learning: A Review of Classification Techniques.” Emerging Artificial Intelligence Applications in Computer Engineering 160: 3–24.

Krantz, David, Duncan Luce, Patrick Suppes, and Amos Tversky. 1971. “Foundations of Measurement, Vol. I: Additive and Polynomial Representations.”

Kroes, Peter. 2002. “Design Methodology and the Nature of Technical Artefacts.” Design Studies, Philosophy of Design, 23 (3): 287–302. https://doi.org/10.1016/S0142-694X(01)00039-4.

Kroes, Peter, Maarten Franssen, Ibo van de Poel, and Maarten Ottens. 2006. “Treating Socio-Technical Systems as Engineering Systems: Some Conceptual Problems.” Systems Research and Behavioral Science 23 (6): 803–814. http://onlinelibrary.wiley.com/doi/10.1002/sres.703/full.

Kuosmanen, Timo. 2008. “Representation Theorem for Convex Nonparametric Least Squares.” The Econometrics Journal 11 (2): 308–325.

Kuosmanen, Timo, and Andrew L. Johnson. 2009. “Data Envelopment Analysis as Nonparametric Least-Squares Regression.” Operations Research 58 (1): 149–60. https://doi.org/10.1287/opre.1090.0722.

Kuosmanen, Timo, Abolfazl Keshvari, and Reza Kazemi Matin. 2015. “Discrete and Integer Valued Inputs and Outputs in Data Envelopment Analysis.” In Data Envelopment Analysis, 67–103. International Series in Operations Research & Management Science. Springer, Boston, MA. https://doi.org/10.1007/978-1-4899-7553-9_4.

Kuosmanen, Timo, and Mika Kortelainen. 2012. “Stochastic Non-Smooth Envelopment of Data: Semi-Parametric Frontier Estimation Subject to Shape Constraints.” Journal of Productivity Analysis 38 (1): 11–28. https://doi.org/10.1007/s11123-010-0201-3.

Latruffe, Laure, Sophia Davidova, and Kelvin Balcombe. 2008. “Application of a Double Bootstrap to Investigation of Determinants of Technical Efficiency of Farms in Central Europe.” Journal of Productivity Analysis 29 (2): 183–91. https://doi.org/10.1007/s11123-007-0074-2.

Laudeman, Irene Vincie, S. G. Shelden, R. Branstrom, and C. L. Brasil. 1998. “Dynamic Density: An Air Traffic Management Metric.” NASA-TM-1998-112226. Moffett Field, California: NASA Ames Research Center. https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19980210764.pdf.

Leveson, Nancy G. 2011. “Applying Systems Thinking to Analyze and Learn from Events.” Safety Science 49 (1): 55–64. http://www.sciencedirect.com/science/article/pii/S0925753510000068.

Lippmann, Richard P. 1987. “Anintroduction to Computing with Neural Nets,”.” IEEE Assp Magazine 4 (2): 4–22.

Liu, John S., Louis Y. Y. Lu, Wen-Min Lu, and Bruce J. Y. Lin. 2013. "A Survey of DEA Applications." Omega 41 (5): 893–902. https://doi.org/10.1016/j.omega.2012.11.004.

Loft, Shayne, Rebekah E. Smith, and Roger W. Remington. 2013. "Minimizing the Disruptive Effects of Prospective Memory in Simulated Air Traffic Control." Journal of Experimental Psychology: Applied 19 (3): 254.

Lovell, CA Knox, Lawrence C. Walters, and Lisa L. Wood. 1994. "Stratified Models of Education Production Using Modified DEA and Regression Analysis." In Data Envelopment Analysis: Theory, Methodology, and Applications, 329–351. Springer.

Maronna, Ricardo A., R. Douglas Martin, Victor J. Yohai, and Matías Salibián-Barrera. 2019. Robust Statistics: Theory and Methods (with R). John Wiley & Sons.

McCulloch, Warren S., and Walter Pitts. 1943. "A Logical Calculus of the Ideas Immanent in Nervous Activity." The Bulletin of Mathematical Biophysics 5 (4): 115–33. https://doi.org/10.1007/BF02478259.

McDonald, John. 2009. "Using Least Squares and Tobit in Second Stage DEA Efficiency Analyses." European Journal of Operational Research 197 (2): 792–98. https://doi.org/10.1016/j.ejor.2008.07.039.

McKinsey. 2019. "Global AI Survey: AI Adoption Proves Its Worth, but Few Scale Impact." https://www.mckinsey.com/featured-insights/artificial-intelligence/global-ai-survey-ai-proves-its-worth-but-few-scale-impact.

Medina-Borja, Alexandra, K. S. Pasupathy, and Konstantinos Triantis. 2007. "Large-Scale Data Envelopment Analysis (DEA) Implementation: A Strategic Performance Management Approach." Journal of the Operational Research Society 58 (8): 1084–1098. http://www.palgrave-journals.com/jors/journal/v58/n8/abs/2602200a.html.

Medina-Borja, Alexandra, and Konstantinos Triantis. 2014. "Modeling Social Services Performance: A Four-Stage DEA Approach to Evaluate Fundraising Efficiency, Capacity Building, Service Quality, and Effectiveness in the Nonprofit Sector." Annals of Operations Research 221 (1): 285–307. https://doi.org/10.1007/s10479-011-0917-0.

Miller, Rupert G. 1974. "The Jackknife--A Review." Biometrika 61 (1): 1. https://doi.org/10.2307/2334280.

Mountrakis, Giorgos, Jungho Im, and Caesar Ogole. 2011. "Support Vector Machines in Remote Sensing: A Review." ISPRS Journal of Photogrammetry and Remote Sensing 66 (3): 247–59. https://doi.org/10.1016/j.isprsjprs.2010.11.001.

Mumford, Enid. 2006. "The Story of Socio-Technical Design: Reflections on Its Successes, Failures and Potential." Information Systems Journal 16 (4): 317–42. https://doi.org/10.1111/j.1365-2575.2006.00221.x.

National Transportation Safety Board. 2016. "Amtrak Train Collision with Maintenance-of-Way Equipment, Chester, Pennsylvania, April 3, 2016." Accident Report NTSB/RAR-17/02. Washington DC: NTSB.

Nørgård, Peter Magnus, Ole Ravn, Niels Kjølstad Poulsen, and Lars Kai Hansen. 2000. "Neural Networks for Modelling and Control of Dynamic Systems-A Practitioner's Handbook."

O'Donnell, Christopher J., D. S. Prasada Rao, and George E. Battese. 2008. "Metafrontier Frameworks for the Study of Firm-Level Efficiencies and Technology Ratios." Empirical Economics 34 (2): 231–55. https://doi.org/10.1007/s00181-007-0119-4.

Olesen, O. B., and N. C. Petersen. 2009. "Target and Technical Efficiency in DEA: Controlling for Environmental Characteristics." Journal of Productivity Analysis 32 (1): 27–40. https://doi.org/10.1007/s11123-009-0133-y.

Oshiro, Thais Mayumi, Pedro Santoro Perez, and José Augusto Baranauskas. 2012. "How Many Trees in a Random Forest?" In Machine Learning and Data Mining in Pattern Recognition, edited by Petra Perner, 154–68. Lecture Notes in Computer Science. Springer Berlin Heidelberg.

Oster, Sharon M. 1999. Modern Competitive Analysis. 3rd ed. Oxford University Press.

O'Sullivan, Arthur, and Steven M Sheffrin. 2007. Economics: Principles in Action. Boston, MA: Pearson/Prentice Hall.

Ozbek, Mehmet Egemen, Jesús M. de la Garza, and Konstantinos Triantis. 2009. "Data Envelopment Analysis as a Decision-Making Tool for Transportation Professionals." Journal of Transportation Engineering 135 (11): 822–831. http://ascelibrary.org/doi/abs/10.1061/(ASCE)TE.1943-5436.0000069.

Pachl, Joern. 2002. Railway Operation and Control. Mountlake Terrace, WA: VTD Rail Publishing.

Paltrinieri, Nicola, Louise Comfort, and Genserik Reniers. 2019. "Learning about Risk: Machine Learning for Risk Assessment." Safety Science 118 (October): 475–86. https://doi.org/10.1016/j.ssci.2019.06.001.

Paradi, Joseph C., and Claire Schaffnit. 2004. "Commercial Branch Performance Evaluation and Results Communication in a Canadian Bank—a DEA Application." European Journal of Operational Research 156 (3): 719–35. https://doi.org/10.1016/S0377-2217(03)00108-5.

Paradi, Joseph C., and H. David Sherman. 2014. "Seeking Greater Practitioner and Managerial Use of DEA for Benchmarking." Data Envelopment Analysis Journal 1 (1): 29–55.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." J. Mach. Learn. Res. 12 (November): 2825–2830. http://dl.acm.org/citation.cfm?id=1953048.2078195.

Peyrache, Antonio, Christiern Rose, and Gabriela Sicilia. 2019. "Variable Selection in Data Envelopment Analysis." European Journal of Operational Research.

Probst, Tahira M., and Ty L. Brubaker. 2001. "The Effects of Job Insecurity on Employee Safety Outcomes: Cross-Sectional and Longitudinal Explorations." Journal of Occupational Health Psychology 6 (2): 139.

Rasmussen, Jens. 1997. "Risk Management in a Dynamic Society: A Modelling Problem." Safety Science 27 (2): 183–213. https://doi.org/10.1016/S0925-7535(97)00052-0.

Ray, Subhash C. 1988. "Data Envelopment Analysis, Nondiscretionary Inputs and Efficiency: An Alternative Interpretation." Socio-Economic Planning Sciences 22 (4): 167–76. https://doi.org/10.1016/0038-0121(88)90003-1.

Rich, Elaine. 1983. "Users Are Individuals: Individualizing User Models." International Journal of Man-Machine Studies 18 (3): 199–214. https://doi.org/10.1016/S0020-7373(83)80007-8.

Roets, Bart, and Johan Christiaens. 2015a. "Evaluation of Railway Traffic Control Efficiency and Its Determinants." European Journal of Transport & Infrastructure Research 15 (4).

Roets, Bart, and Johan Christiaens. 2017. "Shift Work, Fatigue and Human Error: An Empirical Analysis of Railway Traffic Control." Journal of Transportation Safety & Security 0 (ja): 1–18. https://doi.org/10.1080/19439962.2017.1376022.

Roets, Bart, Marijn Verschelde, and Johan Christiaens. 2018. "Multi-Output Efficiency and Operational Safety: An Analysis of Railway Traffic Control Centre Performance." European Journal of Operational Research. https://doi.org/10.1016/j.ejor.2018.04.045.

Rosa, Roger R. 1995. "Extended Workshifts and Excessive Fatigue." Journal of Sleep Research 4 (December): 51–56. https://doi.org/10.1111/j.1365-2869.1995.tb00227.x.

Rosenblatt, F. 1958. "THE PERCEPTRON: A PROBABILISTIC MODEL FOR INFORMATION STORAGE AND ORGANIZATION IN THE BRAIN 1." Psychological Review. Vol. 65.

Rubinstein, Joshua S., David E. Meyer, and Jeffrey E. Evans. 2001. "Executive Control of Cognitive Processes in Task Switching." Journal of Experimental Psychology: Human Perception and Performance 27 (4): 763.

Ruggiero, John. 1996. "On the Measurement of Technical Efficiency in the Public Sector." European Journal of Operational Research 90 (3): 553–65. https://doi.org/10.1016/0377-2217(94)00346-7.

Salmon, Paul M., Neville A. Stanton, Guy H. Walker, Daniel Jenkins, Darshna Ladva, Laura Rafferty, and Mark Young. 2009. "Measuring Situation Awareness in Complex Systems: Comparison of Measures Study." International Journal of Industrial Ergonomics 39 (3): 490–500. https://doi.org/10.1016/j.ergon.2008.10.010.

Salmon, Paul M., Guy H. Walker, and Neville A. Stanton. 2016. "Pilot Error versus Sociotechnical Systems Failure: A Distributed Situation Awareness Analysis of Air France 447." Theoretical Issues in Ergonomics Science 17 (1): 64–79. https://doi.org/10.1080/1463922X.2015.1106618.

Samoilenko, Sergey, and Kweku-Muata Osei-Bryson. 2013. "Using Data Envelopment Analysis (DEA) for Monitoring Efficiency-Based Performance of Productivity-Driven Organizations: Design and Implementation of a Decision Support System." Omega, Data Envelopment Analysis: The Research Frontier - This Special Issue is dedicated to the memory of William W. Cooper 1914-2012, 41 (1): 131–42. https://doi.org/10.1016/j.omega.2011.02.010.

Samuelson, Paul A. 1948. "Consumption Theory in Terms of Revealed Preference." Economica 15 (60): 243–53. https://doi.org/10.2307/2549561.

Santin, Daniel, Francisco J. Delgado, and Aurelia Valino. 2004. "The Measurement of Technical Efficiency: A Neural Network Approach." Applied Economics 36 (6): 627–635.

Scott, W. Richard. 2015. Organizations and Organizing: Rational, Natural and Open Systems Perspectives. Routledge.

Seaver, Bill L., and Konstantinos P. Triantis. 1992. "A Fuzzy Clustering Approach Used in Evaluating Technical Efficiency Measures in Manufacturing." Journal of Productivity Analysis 3 (4): 337–363.

Sheridan, Robert P., Wei Min Wang, Andy Liaw, Junshui Ma, and Eric M. Gifford. 2016. "Extreme Gradient Boosting as a Method for Quantitative Structure–Activity Relationships." Journal of Chemical Information and Modeling 56 (12): 2353–60. https://doi.org/10.1021/acs.jcim.6b00591.

Simar, Léopold, and Paul W. Wilson. 2007. "Estimation and Inference in Two-Stage, Semi-Parametric Models of Production Processes." Journal of Econometrics 136 (1): 31–64. https://doi.org/10.1016/j.jeconom.2005.07.009.

Simar, Léopold, and Paul W. Wilson. 2011. "Two-Stage DEA: Caveat Emptor." Journal of Productivity Analysis 36 (2): 205. https://doi.org/10.1007/s11123-011-0230-6.

Simar, Léopold, and Paul W. Wilson. 2015. "Statistical Approaches for Non-Parametric Frontier Models: A Guided Tour." International Statistical Review 83 (1): 77–110.

Sohn, So Young, and Tae Hee Moon. 2004. "Decision Tree Based on Data Envelopment Analysis for Effective Technology Commercialization." Expert Systems with Applications 26 (2): 279–84. https://doi.org/10.1016/j.eswa.2003.09.011.

Song, Ma-Lin, Ron Fisher, Jian-Lin Wang, and Lian-Biao Cui. 2018. "Environmental Performance Evaluation with Big Data: Theories and Methods." Annals of Operations Research 270 (1): 459–72. https://doi.org/10.1007/s10479-016-2158-8.

Stanton, Kenneth R. 2002. "Trends in Relationship Lending and Factors Affecting Relationship Lending Efficiency." Journal of Banking & Finance 26 (1): 127–152.

Strobl, Carolin, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. 2007. "Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution." BMC Bioinformatics 8 (1): 25. https://doi.org/10.1186/1471-2105-8-25.

Sussman, Donald, and Michael Coplen. 2000. "Fatigue and Alertness in the United States Railroad Industry Part I: The Nature of the Problem." Transportation Research Part F: Traffic Psychology and Behaviour, Fatigue and Transport (I), 3 (4): 211–20. https://doi.org/10.1016/S1369-8478(01)00005-5.

Tan, Pang-Ning. 2018. Introduction to Data Mining. Pearson Education India.

The Aircraft Accident Investigation Bureau of Ethiopia. 2019. "Preliminary Accident Investigation Report of B737-8 (MAX), Registered ET-AVJ." Accident Report AI-01/19.

Tjahjono, Soerjanto. 2018. “Preliminary Accident Investigation Report of PT. Lion Mentari Airlines Boeing 737-8 (MAX); Registered PK-LQP.” Accident Report KNKT.18.10.35.04. Jakarta, Indonesia: Komite Nasional Keselamatan Transportasi (KNKT).

Topcu, Taylan G., and Bryan L. Mesmer. 2018. “Incorporating End-User Models and Associated Uncertainties to Investigate Multiple Stakeholder Preferences in System Design.” Research in Engineering Design 29 (3): 411–31. https://doi.org/10.1007/s00163-017-0276-1.

Topcu, Taylan G., Konstantinos Triantis, and Bart Roets. 2019. “Estimation of the Workload Boundary in Socio-Technical Infrastructure Management Systems: The Case of Belgian Railroads.” European Journal of Operational Research 278 (1): 314–29. https://doi.org/10.1016/j.ejor.2019.04.009.

Triantis, K. 2015. “Engineering Design and Efficiency Measurement: Issues and Future Research Opportunities.” Data Envelopment Analysis Journal 1 (2): 81–112. http://econpapers.repec.org/RePEc:now:jnldea:103.00000008.

Triantis, Konstantinos, Devang Sarayia, and Bill Seaver. 2010. “Using Multivariate Methods to Incorporate Environmental Variables for Local and Global Efficiency Performance Analysis.” INFOR: Information Systems and Operational Research 48 (1): 39–52. https://doi.org/10.3138/infor.48.1.039.

Tu, Jack V. 1996. “Advantages and Disadvantages of Using Artificial Neural Networks versus Logistic Regression for Predicting Medical Outcomes.” Journal of Clinical Epidemiology 49 (11): 1225–31. https://doi.org/10.1016/S0895-4356(96)00002-9.

Turner, Hugh, Robert Windle, and Martin Dresner. 2004. “North American Containerport Productivity: 1984–1997.” Transportation Research Part E: Logistics and Transportation Review 40 (4): 339–356.

Van Dongen, Hans PA, Greg Maislin, Janet M. Mullington, and David F. Dinges. 2003. “The Cumulative Cost of Additional Wakefulness: Dose-Response Effects on Neurobehavioral Functions and Sleep Physiology from Chronic Sleep Restriction and Total Sleep Deprivation.” Sleep 26 (2): 117–126.

Vaneman, Warren K., and Konstantinos Triantis. 2007. “Evaluating the Productive Efficiency of Dynamical Systems.” IEEE Transactions on Engineering Management, 54 (3): 600–612.

Vapnik, Vladimir, Steven E Golowich, and Alex J Smola. 1997. “Support Vector Method for Function Approximation, Regression Estimation and Signal Processing.” In Advances in Neural Information Processing Systems, 281–87.

Vapnik, Vladimir, and A J Lerner. 1963. “Generalized Portrait Method for Pattern Recognition.” Automation and Remote Control 24 (6): 774–80.

Verschelde, Marijn, and Nicky Rogge. 2012. “An Environment-Adjusted Evaluation of Citizen Satisfaction with Local Police Effectiveness: Evidence from a Conditional Data Envelopment Analysis Approach.” European Journal of Operational Research 223 (1): 214–25. https://doi.org/10.1016/j.ejor.2012.05.044.

Vidoli, Francesco, and Giancarlo Ferrara. 2015. “Analyzing Italian Citrus Sector by Semi-Nonparametric Frontier Efficiency Models.” Empirical Economics 49 (2): 641–58. https://doi.org/10.1007/s00181-014-0879-6.

Visser, Ewart de, and Raja Parasuraman. 2011. “Adaptive Aiding of Human-Robot Teaming: Effects of Imperfect Automation on Performance, Trust, and Workload.” Journal of Cognitive Engineering and Decision Making 5 (2): 209–31. https://doi.org/10.1177/1555343411410160.

Wallace, Linda, Mark Keil, and Arun Rai. 2004. “How Software Project Risk Affects Project Performance: An Investigation of the Dimensions of Risk and an Exploratory Model.” Decision Sciences 35 (2): 289–321. http://onlinelibrary.wiley.com/doi/10.1111/j.00117315.2004.02059.x/full.

Weick, Karl E. 1995. “What Theory Is Not, Theorizing Is.” Administrative Science Quarterly, 385–390. http://www.jstor.org/stable/2393789.

Wilson, John R. 2000. “Fundamentals of Ergonomics in Theory and Practice.” Applied Ergonomics 31 (6): 557–67. https://doi.org/10.1016/S0003-6870(00)00034-X.

Wong, M. Anthony, and Tom Lane. 1983. “A Kth Nearest Neighbour Clustering Procedure.” Journal of the Royal Statistical Society. Series B (Methodological) 45 (3): 362–68. http://www.jstor.org/stable/2345405.

Yu, Ming-Miin, and Erwin T. J. Lin. 2008. “Efficiency and Effectiveness in Railway Performance Using a Multi-Activity Network DEA Model.” Omega 36 (6): 1005–17. https://doi.org/10.1016/j.omega.2007.06.003.

Zhao, Y., K. Triantis, P. Murray-Tuite, and P. Edara. 2011. “Performance Measurement of a Transportation Network with a Downtown Space Reservation System: A Network-DEA Approach.” Transportation Research Part E: Logistics and Transportation Review 47 (6): 1140–1159.

Zhu, Nan, Chuanjin Zhu, Ali Emrouznejad, and Luman Chen. 2018. “A Combined DEA with Machine Learning Algorithms for Measuring the Efficiency of Chineese Listed Manufacturing Plants.” In DEA40: Data Envelopment Analysis and Performance Measurement: Recent Developments, 12–21. Birmingham, UK: Aston Business School.

Zhu, Qingyuan, Jie Wu, and Malin Song. 2018. “Efficiency Evaluation Based on Data Envelopment Analysis in the Big Data Context.” Computers & Operations Research 98: 291–300.

APPENDICES

Appendix A. DEA Fundamentals

A *transformation process* or (shortly process) can be defined as the value generating conversion of limited resources, into a set of outputs that are desirable by the stakeholders. *Efficiency Measurement* is a microeconomic field of study that investigates the production performance of **similar** processes or peers, based on their ability to efficiently convert resources to pursue their objectives. Investigated units are denoted as Decision-Making Units (DMUs), based on the assumption that they are capable of pursuing their objectives by making independent decisions. In common notation, DMUs are usually represented with a black-box as shown in Figure A-1. The frame of the Black-box represents both a physical boundary (e.g. similar branches, peers, etc.) and a temporal boundary in the form of a measurement horizon (e.g., a day, a month, an hour, etc.)

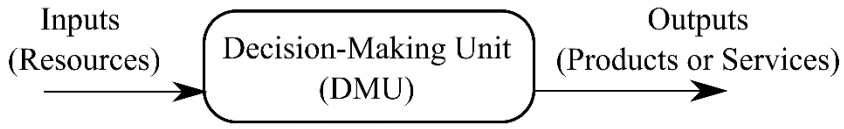

Figure A-1 Generic DEA Black-Box

Similar to the use of the term *efficiency* in engineering and science fields such as thermodynamics; *efficiency* (θ) of a process could be considered as the ratio of outputs to inputs. Like in thermodynamics, this ratio measure is bounded between zero and one; where zero represents pure inefficiency and one represents full technical efficiency (which is not possible to obtain in thermodynamics). Assuming a simple transformation process that consumes a single resource to produce a single output, efficiency could be expressed with the following:

$$0 \leq \text{Efficiency} = \theta = \frac{Output}{Input} \leq 1 \qquad (1)$$

Extension of Equation 1 to a multivariate case could be demonstrated with the following illustrative example that is inspired from Cooper et. al. (2011). Assume that we are evaluating a generic company called *TTECH. TTECH* has nine branches that consume two inputs (*Engineers* and *Designers)* to generate a single output (*Patents)*. For simplicity, assume that all patents, engineers, and designers are of equal value. For a fixed time horizon, lets say a year, we denote branches of *TTECH* with letters and assume that *Table*

A-1 represents their production data, scaled down to its isoquants (unitized to one unit of patent).

Table A-1 Two Inputs - Single Output Case

TTECH	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**
Engineers	4	3	7	4	2	5	6	6	7
Designers	3	6	1	2	4	2	4	3	3
Patents	1	1	1	1	1	1	1	1	1

If we plot the production behavior of TTECH on the axes of Engineers per Patent and Designers per Patent; we would obtain the Production Possibility Set of TTECH that is depicted in Figure A-2.

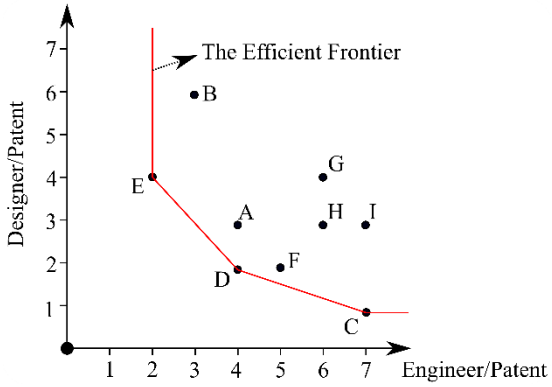

Figure A-2 Production Possibility Set and the Frontier

In Figure A-2, DMUs E, D, and C more convert their resources into desired outputs compared to their peers. In other words, these units are Pareto-Koopmans Efficient units (Koopmans 1951). A DMU is Pareto-Koopmans Efficient (shortly efficient) if and only if it is not possible to improve any input or output without worsening any other input or output. These units are extremely useful to identify because the line connecting these units (the red line in Figure A-2) represents a limit of attainable efficiency or the Pareto-Koopmans frontier (shortly frontier). This frontier can be used as a measure of efficiency for other DMUs in the set, based on their relative distance with respect to it. More specifically, Efficiency of DMU A could be empirically interpreted by drawing a straight line to the origin, and comparing the length of this distance measure to the distance of the frontier to the origin (Farrell 1957). I visualize this idea with the turquoise line in Figure Figure A-3.

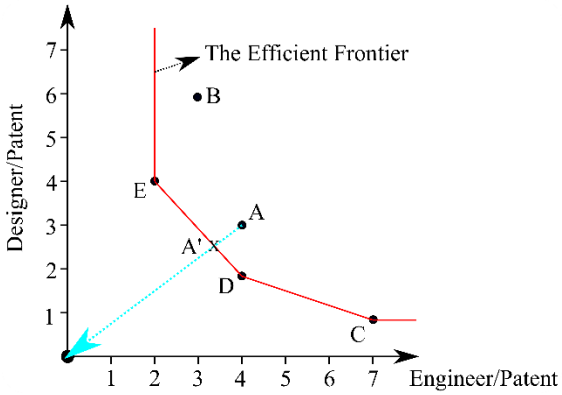

Figure A-3 Radial Efficiency Measure

In Figure A-3, A` represent the point where the radial turquoise distance vector intersects the frontier. Efficiency of DMU A could be estimated by the ratio of the following two distance measures:

$$\text{Efficiency of DMU A} = \frac{|OA`|}{|OA|} \qquad (2)$$

The fundamental assumption here is the following, if DMU A was producing more resources by preserving the same resource consumption, or if it had consumed less resources and produced the same amount, it could have been located anywhere between DMUs E and D, which we roughly represent with A`. This also means that, DMUs D and E are best practices for DMU A to learn from. In efficiency measurement, an inefficient points is evaluated based on its peers, which are defined as the observations that are similar to the evaluated unit yet operate on the frontier. In this case, D and E are peers of A. This also implies that, A` could have obtained by a weighted average of DMU E and DMU D. Therefore, A` is considered a feasible target.

Now that we have established the basics, the notion of efficiency is formally expressed with the following. Given a set of resources $\mathbf{x} = (\mathbf{x_1}, \ldots., \boldsymbol{x_m}) \in \Re_+^m$ that are consumed to produce a set of resources $\boldsymbol{y} = (\mathbf{y_1}, \ldots., \boldsymbol{y_s}) \in \Re_+^s$, the production technology $\boldsymbol{T}$ that maps the inputs to the outputs can be defined as $T(y) = \{x: x\ can\ produce\ y\}, y \in \Re_+^s$. Similarly, a production set Ψ (TTECH in previous example), that includes $\boldsymbol{n}$ number of DMUs, can be defined as:

$$\Psi = \{(x, y) \mid x \in \Re_+^m, \boldsymbol{y} \in \Re_+^s, (x, y)\ is\ feasible\} \qquad (3)$$

Recall that efficient observations are the ones that convert their resources into outputs most efficiently. The formal notation allows to define two virtual weight vectors, v_i and

u_r. These represent the fraction of resource consumption of the evaluated DMU, compared to the efficient observations. We will use these virtual variables to obtain information from the input-output data. More specifically, we will extract how the how Pareto-Koopmans efficient observations used their resources, through the seminal CCR model (Charnes, Cooper, and Rhodes 1978):

To evaluate the efficiency of a DMU_o in our data set (o ranges between 1,2…,n) we need n number of optimization problems, to solve for the values for v_i, (i=1,….,m) and u_r $(r = 1, \dots., s)$:

$$\text{Max}_{\boldsymbol{v},\boldsymbol{u}}\ \theta^* = \frac{u_1 y_{1o} + u_2 y_{2o} + \cdots + u_s y_{so}}{v_1 x_{1o} + v_2 x_{2o} + \cdots + v_m x_{mo}} \quad (4)$$

$$\text{Subject to} \quad \frac{u_1 y_{1j} + \cdots + u_s y_{sj}}{v_1 x_{1j} + \cdots + v_m j} \leq 1, where\ (j = 1, \dots, n) \quad (5)$$

$$\forall v_i \geq 0, and\ \forall u_s \geq 0 \quad (6)$$

Given that X and Y are fixed by data, the objective of Equation 4 is to obtain the values for v_i and u_r that maximize the efficiency score θ. The constraint Equation 5 represent that the ratio of input and output weights should not exceed 1. However, this fractional form is hard to solve, therefore it is linearized so that it can be solved through a linear optimization program. Below, I provide the linearized version of Equation 4-6 (for mathematical proof Charnes, Cooper, and Rhodes 1978):

$$\text{Max}_{\mu,\boldsymbol{v}}\ \theta^* = \mu_1 y_{1o} + \cdots + \mu_s y_{so} \quad (7)$$

$$\text{Subject to} \quad v_1 x_{1o} + \cdots + v_m x_{mo} = 1 \quad (8)$$

$$\mu_1 y_{1j} + \cdots + \mu_s y_{sj} \leq v_1 x_{1j} + \cdots + v_m x_{mj}\ \ where\ (j = 1, \dots, n) \quad (9)$$

$$\forall v_i \geq 0, and\ \forall \mu_s \geq 0 \quad (10)$$

Notice that constraint 6 needs to be imposed n times (for each DMU). Now, we can revisit TTECH data that was provided in Table A1, and solve Equations (7-10) for DMU A. Using TTECH data yields the following:

$$\text{For DMU A} \quad \text{Max}_{\mu,\boldsymbol{v}}\ \theta^A = u \quad (11)$$

$$\text{Subject to} \quad 4 * v_1 + 3 * v_2 = 1 \quad (12)$$

$$\begin{aligned} u &\leq 4 * v_1 + 3 * v_2\ for\ DMU\ A \\ u &\leq 3 * v_1 + 6 * v_2\ for\ DMU\ B \\ u &\leq 7 * v_1 + 1 * v_2\ for\ DMU\ C \\ u &\leq 4 * v_1 + 2 * v_2\ for\ DMU\ D \end{aligned} \quad (13)$$

$$u \leq 2 * v_1 + 4 * v_2 \; for \; DMU \; E$$
$$u \leq 5 * v_1 + 2 * v_2 \; for \; DMU \; F$$
$$u \leq 6 * v_1 + 4 * v_2 \; for \; DMU \; G$$
$$u \leq 6 * v_1 + 3 * v_2 \; for \; DMU \; H$$
$$u \leq 7 * v_1 + 3 * v_2 \; for \; DMU \; I$$
$$\forall v_i \geq 0, and \; \forall \mu_s \geq 0 \qquad (14)$$

Solving the set of linear equations, we observe that the efficiency score for DMU A (θ^A) is equal to 0.857. Our calculations also validate that DMU A's peers are DMUs D and E, as previously indicated by Figure A-2. The relative weights for creating the hypothetical DMU A* are computed as 0.714 and 0.285. The efficiency scores for TTECH companies, computed through the basic CCR model provided in Table A-2.

Table A-2 TTECH Efficiency Scores

TTECH	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**
Engineers	4	3	7	4	2	5	6	6	7
Designers	3	6	1	2	4	2	4	3	3
Patents	1	1	1	1	1	1	1	1	1
CCR Efficiency Scores	0.857	0.667	1.00	1.00	1.00	0.909	0.6	0.667	0.625

In the introduction section, I argued that efficiency measurement is an axiomatic method. Below I provide a brief summary of the axioms of economic production using the generic form described in Equation 3. A detailed discussion is provided elsewhere (Shephard 1970; Färe and Grosskopf 1996).

Axiom 1: No Free Lunch:

$$(x, y) \notin \Psi \text{ if } \mathrm{x} = 0, \mathrm{y} \geq 0, \mathrm{y} \neq 0 \qquad (15)$$

Put simply, Axiom 1 argues that, for all members of Ψ, any input vector could produce zero outputs. Inactivity is always possible.

Axiom 2: Free Disposability:

$$\forall (x, y) \in \Psi, \text{ if } \mathrm{x}' \geq \mathrm{x} \text{ and } \mathrm{y}' \leq \mathrm{y} \text{ then } (x', y') \in \Psi \qquad (16)$$

Axiom 2 simply states that it is always possible to waste resources without obtaining additional outputs.

Axiom 3: Bounded:

$$\forall x \in \Re_{+}^{m} \quad (17)$$

Positive reel numbers bound the set of resources.

Axiom 4: Closedness:

$$\Psi \text{ is closed}$$

Axiom 5: Convexity:

$$\text{if } (x_1, y_1), (x_2, y_2) \in \Psi$$

$$\text{then } \forall \alpha \in [0,1],\ (x, y) = \alpha(x_1, y_1) + (1 - \alpha)(x_2, y_2) \in \Psi \quad (18)$$

This axiom simply states that, weighted average of input output pairs are also members of the production possibility set. This axiom simply enables the approach visualized depicted in Figure A-2.

References

Charnes, Abraham, William W. Cooper, and Edwardo Rhodes. 1978. "Measuring the Efficiency of Decision Making Units." European Journal of Operational Research 2 (6): 429–444. http://www.sciencedirect.com/science/article/pii/0377221778901388.

Cooper, William W., Lawrence M. Seiford, and Joe Zhu. 2011. Handbook on Data Envelopment Analysis. Vol. 164. Springer Science & Business Media.

Färe, Rolf, and Shawna Grosskopf. 1996. "Static Production Structure." In Intertemporal Production Frontiers: With Dynamic DEA, edited by Rolf Färe and Shawna Grosskopf, 9–45. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-009-1816-0_2.

Farrell, M. J. 1957. "The Measurement of Productive Efficiency." Journal of the Royal Statistical Society. Series A (General) 120 (3): 253–90. https://doi.org/10.2307/2343100.

Koopmans, Tjalling C. 1951. An Analysis of Production as an Efficient Combination of Activities. Cowles Commission for Research in Economics. New York: John Wiley & Sons.

Shephard, R.W. 1970. Theory of Cost and Production Function. Princeton, NJ: Princeton University Press.

Appendix B. Variables Included in the Model

The definitions of variables that are included in the measurement framework are as follows:

Input (X) Variables:

Controller Skill Level: We measure skill levels with the number days INFRABEL assigns the Controller to the designated role. We use this variable as an input variable for both Controllers based on the understanding that the primary resource for the transformation process in this study is one hour of that Controller's time.

Output (Y) Variables:

Manual Move Decisions: Represents the number of times the TC manually activates a railroad signal to let a train pass. It is composed of various subtypes, however all movement decisions are fundamentally made to move trains through the infrastructure and do not aim to prevent accidents. The unit of measurement is its quantity.

Auto Move Decisions: Represents traffic control decisions that are the same with the type described previously however, the automated traffic control decision-aide tool makes these decisions. The TC can switch the level of automation any time for any desired period during the duration of the shift and the allowed automation range is between 0-100%. The decision-aid tool is capable of handling traffic decisions in non-complex traffic (such as in rural areas). However, Controllers do not use the tool under dense and complex traffic. Its unit of measurement is its quantity.

Adapt Decisions: Represents the number of railroad track-adaptation decisions. These types of decisions require manual intervention by the Controller to change the status of the railroad switches, in other words it modifies the state of the railroad usually to ensure the trains will remain on schedule. Examples include merging or splitting trains, re-routing of trains, or special procedures at single-track lines. Controllers do not make these decisions to prevent an incident. However, both types of Controllers perform these types of decisions. The unit of measurement is the weighted amount of actions. We calculate the weights based on the number of seconds that the Controllers operate their systems.

Safety Decisions: This variable represents the total number of safety interventions made by SCs. Examples of safety decisions are the protection of track maintenance sites through safety locks in the signaling system, or launching safety procedures at level

crossings. Similar to the Adapt Decisions, the unit of measurement is the weighted amount of actions.

Monitored Traffic: This variable represents the trains travelling through the SCs dedicated control area and we measure it by the number of signal passes by the trains within the network. Different from other output variables, we treat this variable as an uncontrollable variable as the SCs have zero control over the train passes under nominal conditions.

Socio-technical Performance Environment (Z) Variables:

Traffic Density: This variable represents the density of the railroad traffic and we calculate it by dividing the number of train movements with the number of large traffic control signals controlled by that Controller. It differs from traffic complexity, as a highly dense traffic might not be too complex if the trains run in parallel or do not pass delays to each other. Uncertainty associated with this variable is not considered.

Traffic Complexity: We calculate this variable by dividing the number of adaptation decisions with the movements through the area and we consider it as an indicator of the dedicated railroad network portion's traffic complexity. The assumption behind this variable is that as the traffic becomes more complex, the Controllers require more adaptation decisions to prevent conflicts. Uncertainty associated with this variable is not considered.

Safety Complexity: This variable is an attempt to represent the complexity of safety decisions performed by the Safety Controllers. As argued earlier, not all safety decisions are equal and some involve more effort to perform. Besides there are cognitive capability related effects of switching between certain decision tasks (Rubinstein et al., 2001). Therefore, we include this variable while recognizing it is a naïve proxy of complexity. We calculate this variable by dividing the total number of safety decisions made during one-hour measurement period with the total number of possible safety decisions. Uncertainty associated with this variable is not considered.

Fatigue Level: This variable represents the mental fatigue of personnel. We know this to be an important factor associated with human caused errors, as discussed in Section 2. Fatigue risk level used in this paper is calculated by INFRABEL's predictive tool that is conceptually based on the fatigue Risk Index (Folkard et al., 2007) and is validated for the

railway traffic environment at INFRABEL. Fundamentally, the predictive tool is built on the statistical relationship between the human induced error and work shift trends of Controllers in the TCCs, and its fundamentals are described elsewhere (Roets & Christiaens, 2017). We calculate the fatigue level for the entire duration of the shift for each Controller and is a maximum likelihood estimator of how much more likely an average Controller is expected to make mistakes under physical distress, compared to a standard staff schedule, which receives an estimated fatigue risk level of 1. This variable is unit-less and we observe the range of the variable to be 0.5 and 1.5. There is some built in uncertainty associated with this variable that we associate with the predictive nature of the model. Human fatigue is a complex phenomenon, and the fatigue measure in this study is only a proxy.