



Poznan University of Technology
Faculty of Computing and Telecommunications

Doctoral dissertation

New Directions
in Robust Data Envelopment Analysis

Anna Labijak-Kowalska

Supervisor: Miłosz Kadziński, Ph.D., Habil., Assoc. Prof.

Poznań, 2023

Abstract

Data Envelopment Analysis (DEA) is a method for evaluating the efficiency of the Decision Making Units (DMUs) consuming multiple inputs and producing multiple outputs. Its applications include real-world problems from various fields, e.g., banking, healthcare, education, or transportation. In the original DEA setting, the performance of a unit is assessed using the single, most favorable input-output weight vector. It allows only to indicate the set of efficient DMUs without the possibility of comparing them. In this dissertation, we address these drawbacks and propose a novel framework for robustness analysis in the context of DEA. First, we propose a set of methods that explore the complete spectrum of feasible weight vectors using two efficiency models, i.e., the basic ratio-based model and the Value-based additive DEA (VDEA) model inspired by the field of MCDA. The framework introduced in this dissertation consists of two complementary types of methods: the exact ones, based on mathematical programming, and the stochastic ones, based on the Monte Carlo simulation. We consider three different viewpoints of the analysis: the efficiency scores, the pairwise comparisons between units, and the efficiency rankings. In addition, this dissertation extends the proposed approach to consider the imprecise information (in the form of interval and ordinal input/output performances and the admissible ranges of marginal functions) and the hierarchical structure of indicators. We propose the algorithms to determine the efficiency reducts and constructs useful for explaining the recommendations. Moreover, the method for finding the single representative weight vector based on the outcomes of the robustness analysis was proposed. Furthermore, we have introduced some measures which aggregate the outcomes for multiple scenarios. Finally, this dissertation contains the experimental comparison of the existing approaches providing the full ranking of DMUs. The applicability of the methodological contribution of this dissertation is illustrated with some real-world case studies including, among others, the efficiency evaluation of Polish airports, Emergency Department physicians or resilience of countries' electricity systems.

List of publications

The dissertation consists of the introductory section and the following seven original publications:

- [P1] M. Kadziński, A. Labijak, and M. Napieraj. Integrated framework for robustness analysis using ratio-based efficiency model with application to evaluation of polish airports. *Omega*, 67:1–18, 2017, DOI: 10.1016/j.omega.2016.03.003.

Number of citations¹:

- according to Web of Science: 24
- according to Google Scholar: 39

- [P2] P. Gasser, M. Cinelli, A. Labijak, M. Spada, P. Burgherr, M. Kadziński, and B. Stojadinović. Quantifying electricity supply resilience of countries with robust efficiency analysis. *Energies*, 13(7):1535, 2020, DOI: 10.3390/en13071535.

Number of citations¹:

- according to Web of Science: 7
- according to Google Scholar: 9

- [P3] A. Labijak-Kowalska and M. Kadziński. Experimental comparison of results provided by ranking methods in data envelopment analysis. *Expert Systems with Applications*, 173:114739, 2021, DOI: 10.1016/j.eswa.2021.114739.

Number of citations¹:

- according to Web of Science: 6
- according to Google Scholar: 11

- [P4] A. Labijak-Kowalska, M. Kadziński, I. Spychała, L. C. Dias, J. Fiallos, J. Patrick, W. Michalowski, and K. Farion. Performance evaluation of emergency department physicians using robust value-based additive efficiency model. *International Transactions in Operational Research*, 30(1):503–544, 2023, DOI: 10.1111/itor.13099.

Number of citations¹:

- according to Web of Science: 1
- according to Google Scholar: 3

¹as on May 30, 2023

- [P5] A. Labijak-Kowalska and M. Kadziński. Exact and stochastic methods for robustness analysis in the context of imprecise data envelopment analysis. *Operational Research*, 23(1):22, Mar 2023, DOI: 10.1007/s12351-023-00755-z.

Number of citations¹:

- according to Google Scholar: 1

- [P6] A. Labijak-Kowalska, M. Kadziński, and L. C. Dias. Robustness analysis for imprecise additive value efficiency analysis with an application to evaluation of special economic zones in poland. 2023. Submitted to Socio-Economic Planning Sciences

- [P7] A. Labijak-Kowalska, M. Kadziński, and W. Mrozek. Robust additive value-based efficiency analysis with a hierarchical structure of inputs and outputs. *Applied Sciences*, 13(11), 2023

Contents

1	Introduction	1
2	Data Envelopment Analysis	5
2.1	Ratio-based efficiency model	5
	Output oriented problem	9
2.2	Other Data Envelopment Analysis models	10
	BCC model	10
	Additive model	11
	Value-based additive DEA model	13
2.3	Super-efficiency	14
2.4	Cross-efficiency	16
2.5	Robustness analysis in Data Envelopment Analysis	17
3	Robustness analysis framework for Data Envelopment Analysis	19
3.1	Basic concepts	19
3.2	Efficiency scores	20
3.3	Efficiency ranks	21
3.4	Preference relations	23
3.5	Robustness analysis for Value-based additive efficiency model	24
3.6	Interdependencies between the robust and stochastic results	25
3.7	Evolution of robust results with incremental specification of weight constraints	26
3.8	Robustness analysis for Data Envelopment Analysis with imprecise information	27
	Mathematical programming methods	27
	Stochastic methods	31
3.9	Robustness analysis for Data Envelopment Analysis with a hierarchical structure of inputs and outputs	31
	Mathematical programming methods	32
	Simulation-based methods	34
3.10	Robustness analysis methods for multiple scenarios of efficiency evaluation	34
4	Extensions and applications	37
4.1	Selection of a common vector of weights based on the outcomes of robustness analysis	37
4.2	Efficiency reducts and coconstructs	38
4.3	Experimental comparison of ranking methods in Data Envelopment Analysis	39

4.4 Case studies	43
5 Summary	47
Bibliography	51
Publication reprints	59
Extended abstract in Polish	287
Declarations	301

Chapter 1

Introduction

Data Envelopment Analysis (DEA) is a method allowing to evaluate the relative efficiency of Decision Making Units (DMUs). A DMU can be any entity that consumes multiple inputs and produces multiple outputs. The idea of DEA has its origins in Farrell's definition of performance [33]. He proposed a method to assess the efficiency of a unit using a frontier production function, in contrast to the average production used in most of the literature up to this time. His idea, focused on a single-output scenario, was further developed, resulting in the work [16], which forms the origin of DEA. In their problem formulation, the efficiency of a unit was defined as a ratio between a virtual output and a virtual input. The goal of the standard DEA method is to identify the set of units that perform efficiently. It is done by finding the single vector of input and output weights, which is the most favorable for the examined unit, i.e., its efficiency score is the greatest possible.

During the last 50 years, multiple researchers explored the DEA approach resulting in new models of efficiency and multiple extensions [20, 31] and applications. The most popular areas of DEA applications are banking, health care, agriculture, transportation, and education [60, 77]. The DEA-related literature contains over ten thousand journal papers with continuous and rapid growth [32].

The main DEA models are the CCR model [16], the BCC model [10], and the additive model [15]. The first two are the radial models and require distinguishing between the input and output orientation. The latter is a non-radial model, in which the efficiency score is based on the L_1 distance to the efficient frontier. Such an approach combines both orientations into a single model [23]. Such a combination is often desirable. However, the original formulation of the additive model has some drawbacks. First, there is a scale problem – the projections of the inefficient units on the efficient frontier depend significantly on the scale of inputs and outputs. Moreover, the efficiency measure in this model has no intuitive interpretation [38]. To address these problems, in [26, 38], the authors proposed another additive model for DEA, which draws from the field of Multi-Criteria Decision Analysis (MCDA). They transform the inputs and outputs into possibly non-linear value functions and aggregate them using an additive function based on the concept of the Multi-Attribute Value Theory (MAVT) [34, 48].

Independently on the chosen model, the evaluation of DMUs is always based only on the most favorable scenario for the examined DMU. The vectors of weights assigned to inputs and outputs are different for each unit, which casts doubts about the legitimacy

of the comparisons between them [58]. Moreover, the standard DEA setting does not respond to the question of how the DMUs perform under other weight vectors [70]. Moreover, the efficiency scores are highly sensitive to changes in the set of DMUs under consideration [71, 89]. Finally, DEA allows only to indicate the set of efficient DMUs, not providing tools to discriminate between them.

These drawbacks were an area of concern for further research. First, multiple methods that allow discrimination between the efficient units and provide the complete ranking of DMUs were proposed [1, 3]. The most popular among them are cross-efficiency [72] and super-efficiency [9]. Second, the preference information was included to limit the feasible weight vector space [65, 79]. The above techniques address some of the problems related to the original DEA. However, they do not explore the whole spectrum of weights.

In this dissertation, the novel framework for the robustness analysis in a context of DEA was proposed for both the ratio-based and the value-based efficiency models. It provides a set of methods that explore the whole spectrum of feasible weight vectors. We consider three different points of view: the efficiency scores (or distances to the best unit), efficiency ranks, and the pairwise comparisons between the DMUs. On the one hand, for each of these viewpoints, we propose the methods based on the mathematical programming providing the extreme (minimal and maximal) values for the considered measure, the extreme efficiency scores, extreme distances to the best DMU, and extreme efficiency ranks. For the pairwise comparisons perspective, we define the necessary and possible efficiency preferences for pairs of DMUs and verify the truth of such relations. On the other hand, the intervals between the extreme values are, in many cases, very wide, so the robustness analysis framework incorporates also some stochastic methods, based on Monte Carlo simulation, allowing us to determine the distributions of considered metrics over the feasible weight vector space. We capture these distributions by acceptability indices, such as the Efficiency Acceptability Interval Indices (EAIIs), Distance Acceptability Interval Indices (DAIIs), Efficiency Rank Acceptability Indices (ERAIIs) and Pairwise Efficiency Outranking Indices (PEOIs).

In many real-world problems, the provided information is imperfect. Sometimes, the input and output values cannot be measured precisely, or such a measurement would be too expensive. It is also possible that these values change over time [8, 25]. This is why the robustness analysis methods introduced in this dissertation were adapted for the situation where inputs and outputs are imprecisely defined. In a context of DEA, two types of uncertainty are usually considered, the interval and ordinal factors [22, 60]. Both of them are included in this dissertation. Moreover, for problems where the Value-based additive DEA (VDEA) model is used, we also consider a third type of imprecise information concerning the marginal value functions. The precise value functions may be replaced with the range of admissible values defined by two boundary functions.

In the standard DEA applications, the structure of inputs and outputs is flat. Often, it is beneficial for the users to organize them in a hierarchical structure of categories. The higher-lever factors represent more general concepts, while the lower-lever are more specific. Such a structure has some useful properties [24]. First, modifying and updating the hierarchy is easy if new information becomes available. Second, it allows to decompose the problem into smaller pieces, which are more manageable and provide more specific information. Third, with the hierarchical structure, it is possible to model the interactions between the categories, not only the individual factors. The weight constraints may be defined at any level of the hierarchy. In MCDA, the benefits of the hierarchical

structure were widely explored (e.g., [24, 27, 68]). In a context of DEA, the Multiple-Layer DEA (MLDEA) model was proposed to handle an arbitrary number of levels [73]. Moreover, in this work, the inputs and outputs must be organized into two separate hierarchies. In this dissertation, we introduce the model which uses the VDEA to evaluate the DMUs with inputs and outputs organized into a single hierarchy. We also adapt the robustness analysis framework to the hierarchical DEA problems.

The outcomes of the robustness analysis framework may be, sometimes, hard to understand and interpret, so we introduced the method which provides a single weight vector, which represents the most the outcomes of the robustness analysis. Based on this vector, one may evaluate the whole set of units and generate their ranking with an advantage of a common base while retaining the idea of representation of the whole feasible weight vector space.

The last two extensions of the proposed framework include the multiple scenario analysis, which is applicable in situations when the same set of DMUs is evaluated under different scenarios. We propose some methods which aggregate the outcomes from individual scenarios. Moreover, we define and propose the algorithm to determine the efficiency reducts and constructs. For the efficient units, we indicate the efficiency reducts, i.e., the minimal subsets of inputs and outputs, for which the given unit is efficient. For inefficient DMUs, we search for the efficiency constructs, i.e., the minimal subsets of units, which should be removed from the data set to make the examined unit efficient.

The methodological part of this dissertation is illustrated with real-world case studies. First, the robustness analysis framework with the ratio-based efficiency model is applied to evaluate Polish airports and the electricity supply resilience of different countries. In the latter, we applied also the algorithm for the determination of the efficiency reducts and constructs. Second, we analyzed the 20 Emergency Department physicians using the VDEA model enriched with identification of the representative set of weights and the multiple-scenario analysis for different patients' complaint groups. Third, several case studies, including the evaluation of Special Economic Zones in Poland, Chinese ports, and industrial robots, were conducted using the robustness analysis for DEA with imprecise data. Finally, the case study of the healthcare systems in Polish voivodeships was considered using the VDEA with hierarchical structure of factors.

Furthermore, this dissertation includes the experimental comparison of the methods providing the full ranking of DMUs in a context of DEA. We identified and implemented fifteen ranking procedures, which represent different concepts, such as super- and cross-efficiency, multivariate statistics, the role of units as benchmarks for others, and the outcomes of the robustness analysis. The experiments included ten real-world and 960 randomly generated data sets. The rankings provided by the analyzed methods were compared using five measures: hit ratio and normalized hit ratio for the choice problem, as well as Kendall's τ , Rank Difference Measure (RDM), and Rank Acceptance Measure (RAM) for complete rankings [47]. The obtained results allowed us to identify some groups of procedures for which the provided rankings are similar and some procedures representing unique concepts. Such an analysis, combined with identifying the strengths and weaknesses of each procedure, eases the users to choose the method that is the most suitable for their problem.

The remainder of this doctoral dissertation is organized in the following way. Chapter 2 discusses the DEA method, its main efficiency models, concepts of the super- and

cross-efficiency, and the existing robustness analysis approaches. Chapter 3 extensively describes the proposed robustness analysis framework for the ratio-based and VDEA efficiency models. In Chapter 4, we present the extensions of the proposed framework and the conducted case studies. Chapter 5 concludes the dissertation and presents some areas for future research.

Chapter 2

Data Envelopment Analysis

This chapter describes the idea of DEA. We focus on the main efficiency models and delineate the general aim of the traditional efficiency analysis.

2.1 Ratio-based efficiency model

This section describes the original ratio-based efficiency model of DEA [16]. It is inspired by the commonly used measure of efficiency for the DMUs with one input and one output:

$$\text{efficiency} = \frac{\text{output}}{\text{input}}. \quad (2.1)$$

For example, when evaluating an employee, the managers can determine the “output per hour”, while for the production companies, the performance is usually calculated with the “cost per unit” measure. Such an approach is intuitive. However, it does not deal with the situation when the units must be evaluated in terms of consuming multiple inputs and producing multiple outputs.

We consider a set of K DMUs, denoted as $\mathcal{D} = \{DMU_1, DMU_2, \dots, DMU_K\}$. Each unit is described with a set of M inputs ($\mathcal{IN} = \{x_1, x_2, \dots, x_M\}$) and a set of N outputs ($\mathcal{OUT} = \{y_1, y_2, \dots, y_N\}$). The values of k -th DMU on m -th input and n -th output are denoted as, respectively, $x_{m,k}$ and $y_{n,k}$.

The efficiency score of DMU_o is defined as the ratio of the virtual output and the virtual input:

$$E_o = \frac{\sum_{n=1}^N \mu_n y_{n,o}}{\sum_{m=1}^M \nu_m x_{m,o}}, \quad (2.2)$$

where μ_m and ν_n are, respectively, weights assigned to the m -th input and n -th output.

To evaluate the efficiency of a given unit DMU_o , one may want to determine the most favorable input-output weight vector for this DMU. Such the most favorable weight vector can be determined by solving the following mathematical programming model:

$$\begin{aligned} \max \quad & E_o = \frac{\sum_{n=1}^N \mu_n y_{n,o}}{\sum_{m=1}^M \nu_m x_{m,o}} \\ \text{subject to:} \quad & \frac{\sum_{n=1}^N \mu_n y_{n,k}}{\sum_{m=1}^M \nu_m x_{m,k}} \leq 1, & \text{for } k = 1, 2, \dots, K, \\ & \mu_n, \nu_m \geq 0, & \text{for } n = 1, 2, \dots, N, \quad m = 1, 2, \dots, M. \end{aligned} \quad (2.3)$$

In this model, the ratio-based efficiency score for considered DMU is maximized with the restriction that the efficiency scores do not exceed 1 for all units in the data set.

As the model above is not linear, the authors used the Charnes-Cooper transformation [14] to obtain the equivalent linear model:

$$\begin{aligned}
\max \quad & E_o = \sum_{n=1}^N \mu_n y_{n,o} \\
\text{subject to:} \quad & \sum_{m=1}^M \nu_m x_{m,o} = 1, \\
& \sum_{n=1}^N \mu_n y_{n,k} \leq \sum_{m=1}^M \nu_m x_{m,k}, \quad \text{for } k = 1, 2, \dots, K, \\
& \mu_n, \nu_m \geq 0, \quad \text{for } n = 1, 2, \dots, N, \quad m = 1, 2, \dots, M.
\end{aligned} \tag{2.4}$$

In this model, the weighted sum of outputs for analyzed DMU is maximized with the restriction that the weighted sum of inputs for this unit is equal to 1. The remaining constraints ensure that the efficiency score of all DMUs is not greater than 1 and that the inputs' and outputs' weights are non-negative. If the optimal objective value (E_o^*) is equal to 1 for some DMU_o , it means that there exists at least one input-output weight vector, for which DMU_o is the best unit in terms of the ratio-based efficiency score. In this case, DMU_o is deemed weakly efficient. If, in addition, there exists at least one solution where all input and output weights are strictly positive, then DMU_o is called CCR efficient. The opposite situation ($E_o^* < 1$) means that, for all weight vectors, there is at least one other DMU better than DMU_o , so DMU_o is inefficient.

Example For illustrative purposes, let us consider six DMUs described with single input and single output (see Table 2.1). The efficiency of DMU E can be determined by solving the model below:

$$\begin{aligned}
\max \quad & E_E = 4\mu_{out} \\
\text{subject to:} \quad & 2.5\nu_{in} = 1, \\
& 6\mu_{out} \leq 2\nu_{in} \quad (A), \\
& 2\mu_{out} \leq \nu_{in} \quad (B), \\
& 7\mu_{out} \leq 3\nu_{in} \quad (C), \\
& 6\mu_{out} \leq 4\nu_{in} \quad (D), \\
& 4\mu_{out} \leq 2.5\nu_{in} \quad (E), \\
& 2.5\mu_{out} \leq 3.5\nu_{in} \quad (F).
\end{aligned} \tag{2.5}$$

Based on the optimal solution of the model above, the CCR efficiency score of E is equal to 0.533, so this unit is inefficient. The input and output weights are equal, respectively, $\nu_{in} = 0.4$ and $\mu_{out} = 0.133$. Similarly, the efficiency score for other units can be determined.

The CCR DEA model also takes the alternative form, which is a dual formulation of the problem presented in Equation 2.4. In this case, the model does not focus on the efficiency scores of DMUs but on the efficient frontier formed by a linear combination

Table 2.1: Input and output values for DMUs from the considered example.

DMU	input	output
A	2	6
B	1	2
C	3	7
D	4	6
E	2.5	4
F	3.5	2.5

of efficient units. The Linear Programming model for the dual formulation of the CCR model is presented below:

$$\begin{aligned}
 & \min \theta \\
 \text{subject to: } & \sum_{k=1}^K \lambda_k x_{m,k} \leq \theta x_{m,o}, \quad \text{for } m = 1, 2, \dots, M, \\
 & \sum_{k=1}^K \lambda_k y_{n,k} \geq y_{n,o}, \quad \text{for } n = 1, 2, \dots, N, \\
 & \lambda_k \geq 0, \quad \text{for } k = 1, 2, \dots, K, \\
 & \theta \geq 0.
 \end{aligned} \tag{2.6}$$

In the dual formulation of the CCR model, we search for a virtual unit (a conical combination of the existing DMUs), which attains better performance, i.e., lower inputs and greater outputs than the analyzed DMU_o . Moreover, the input values of DMU_o are multiplied with a variable θ , which is minimized. This assures that all inputs are reduced proportionally (radially) as most as possible – the virtual DMU is as far from DMU_o as possible. If the optimal value of θ (θ^*) is equal to 1, then the DMU_o is compared to itself ($\lambda_o = 1, \lambda_k = 0$ for $k = 1, 2, \dots, K, k \neq o$). This situation takes place for units deemed as weakly efficient. For inefficient DMUs θ^* is lesser than 1.

Example When considering unit E from the exemplary data set, the dual CCR model takes the following form:

$$\begin{aligned}
 & \min \theta \\
 \text{subject to: } & 2\lambda_A + \lambda_B + 3\lambda_C + 4\lambda_D + 2.5\lambda_E + 3.5\lambda_F \leq 2.5\theta \\
 & 6\lambda_A + 2\lambda_B + 7\lambda_C + 6\lambda_D + 4\lambda_E + 2.5\lambda_F \geq 4 \\
 & \lambda_A, \lambda_B, \lambda_C, \lambda_D, \lambda_E, \lambda_F \geq 0, \\
 & \theta \geq 0.
 \end{aligned} \tag{2.7}$$

To determine if some weakly efficient unit is fully efficient (CCR efficient), we need

to solve the following LP model, which maximizes the slacks for a considered DMU:

$$\begin{aligned}
& \max \sum_{m=1}^M s_m^- + \sum_{n=1}^N s_n^+ \\
& \text{subject to: } \sum_{k=1}^K \lambda_k x_{m,k} = \theta^* x_{m,o} - s_m^-, & \text{for } m = 1, 2, \dots, M, \\
& \sum_{k=1}^K \lambda_k y_{n,k} = y_{n,o} + s_n^+, & \text{for } n = 1, 2, \dots, N, \\
& \lambda_k \geq 0, & \text{for } k = 1, 2, \dots, K, \\
& s_m^- \geq 0, s_n^+ \geq 0, & m = 1, 2, \dots, M, n = 1, 2, \dots, N,
\end{aligned} \tag{2.8}$$

where θ^* is an optimal solution for the dual CCR model (Equation 2.6). If all slacks in the optimal solution (s_m^- and s_n^+) are zeros, then the considered DMU_o is CCR efficient.

The linear combination of DMUs obtained with the dual formulation of the CCR model forms an artificial DMU which is a projection of the analyzed DMU_o into the efficiency frontier. This artificial unit is called a Hypothetical Comparison Unit (HCU). Its input and output values are calculated as follows:

$$x_{m,HCU} = \sum_{k=1}^K \lambda_k^* x_{m,k}, \tag{2.9}$$

$$y_{m,HCU} = \sum_{k=1}^K \lambda_k^* y_{n,k}, \tag{2.10}$$

where λ_k^* are the optimal values from model 2.6. Figure 2.1 presents, graphically, the efficiency frontier and the HCU for the DMU E from the considered example.

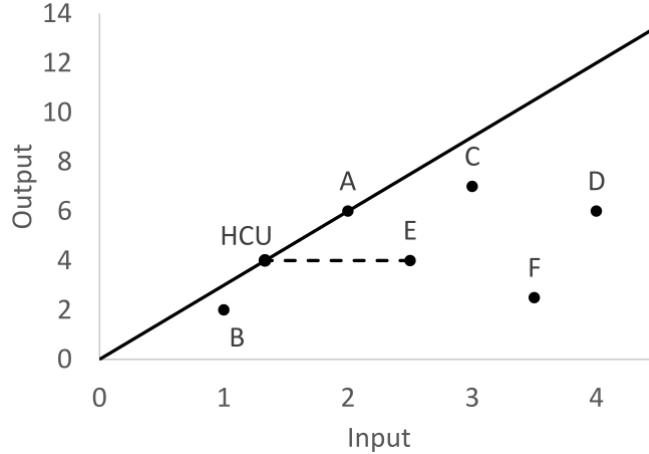


Figure 2.1: Hypothetical Comparison Unit (HCU) for exemplary DMU E .

Having determined the inputs and outputs values of the HCU, it is possible to determine the improvements that need to be implemented for all inputs simultaneously to achieve efficiency. The necessary improvement of m -th input or n -th output can be

obtained as follows:

$$\Delta x_{m,o} = x_{m,o} - x_{m,HCU}, \quad (2.11)$$

$$\Delta y_{n,o} = y_{n,HCU} - y_{n,o}. \quad (2.12)$$

Output oriented problem

Up to this point, we focused on the input-oriented perspective, i.e., the objective was to minimize the inputs and remain the outputs at the same level. Hence, it is called an input-oriented model. However, it is possible to conduct the efficiency analysis from the perspective of increasing outputs with inputs remaining at the same level. Such models are called output-oriented. The primal formulation of the output-oriented efficiency model is the following:

$$\begin{aligned} \min \quad & \frac{1}{E_o} = \sum_{m=1}^M \nu_m x_{m,o} \\ \text{subject to:} \quad & \sum_{n=1}^N \mu_n y_{n,o} = 1, \\ & \sum_{n=1}^N \mu_n y_{n,k} \leq \sum_{m=1}^M \nu_m x_{m,k}, \quad \text{for } k = 1, 2, \dots, K, \\ & \mu_n, \nu_m \geq 0, \quad \text{for } n = 1, 2, \dots, N, \quad m = 1, 2, \dots, M. \end{aligned} \quad (2.13)$$

This time, the weighted sum of outputs is restricted to the value of one, while the weighted sum of inputs for DMU_o is minimized. The normalizing constraints remain the same as in the input-oriented model. The efficiency score of DMU_o is the inverse of the objective value obtained with this model.

The following LP model presents the dual (envelopment) formulation for the output-oriented problem for DMU_o . In this model, the inputs of the artificial DMU need to be at most as big as for DMU_o , while the outputs must be not lesser than these of DMU_o multiplied by a variable θ which is maximized. So, the model searches for the maximal value by which all outputs of a considered unit can be increased proportionally without exceeding the outputs of the artificial DMU.

$$\begin{aligned} \max \quad & \theta \\ \text{subject to:} \quad & \sum_{k=1}^K \lambda_k x_{m,k} \leq x_{m,o}, \quad \text{for } m = 1, 2, \dots, M, \\ & \sum_{k=1}^K \lambda_k y_{n,k} \geq \theta y_{n,o}, \quad \text{for } n = 1, 2, \dots, N, \\ & \lambda_k \geq 0, \quad \text{for } k = 1, 2, \dots, K. \end{aligned} \quad (2.14)$$

Again, the inverse of the optimal objective value (θ^*) is the efficiency score of DMU_o .

Example Both formulations of the output-oriented model allow evaluating the DMU E from the considered example are presented below.

Primal formulation:

$$\begin{aligned}
 & \min 2.5\nu_{in} \\
 & \text{s.t. } 4\mu_{out} = 1, \\
 & \quad 6\mu_{out} \leq 2\nu_{in} \quad (A), \\
 & \quad 2\mu_{out} \leq \nu_{in} \quad (B), \\
 & \quad 7\mu_{out} \leq 3\nu_{in} \quad (C), \\
 & \quad 6\mu_{out} \leq 4\nu_{in} \quad (D), \\
 & \quad 4\mu_{out} \leq 2.5\nu_{in} \quad (E), \\
 & \quad 2.5\mu_{out} \leq 3.5\nu_{in} \quad (F).
 \end{aligned} \tag{2.15}$$

Dual formulation:

$$\begin{aligned}
 & \max \theta \\
 & \text{s.t. } 2\lambda_A + \lambda_B + 3\lambda_C + 4\lambda_D + 2.5\lambda_E + 3.5\lambda_F \leq 2.5, \\
 & \quad 6\lambda_A + 2\lambda_B + 7\lambda_C + 6\lambda_D + 4\lambda_E + 2.5\lambda_F \geq 4\theta, \\
 & \quad \lambda_A, \lambda_B, \lambda_C, \lambda_D, \lambda_E, \lambda_F \geq 0, \\
 & \quad \theta \geq 0
 \end{aligned} \tag{2.16}$$

Similarly to the input-oriented model, the linear combination of DMUs obtained with the dual formulation of DEA model (λ_k) forms a HCU and allows to find the necessary improvements of outputs to become efficient. The projections of all inefficient units onto the efficient frontier in both: input-oriented in output-oriented models are shown in Figure 2.2.

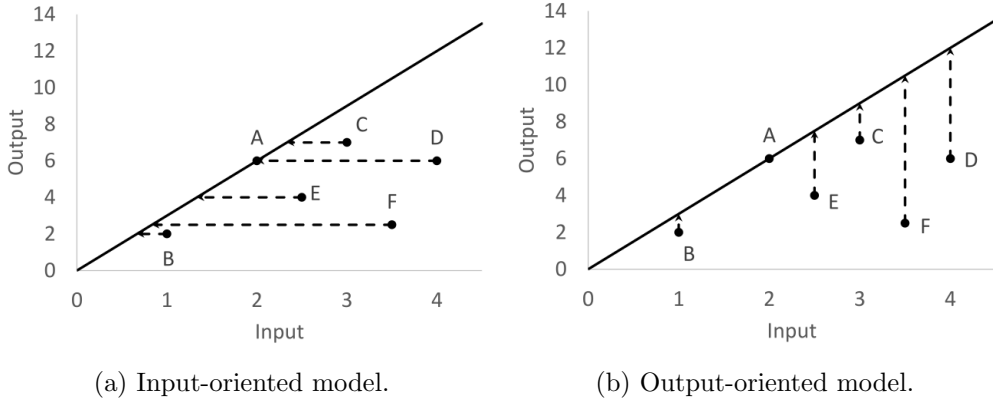


Figure 2.2: Comparison between the HCUs in input and output orientated DEA models.

2.2 Other Data Envelopment Analysis models

In the previous part of this dissertation, we described the original ratio-based efficiency model, called CCR. It is based on the assumption of the constant returns to scale of activities, i.e., it assumes that any production plan, which is a conical combination of the existing units, is feasible. Such assumption can be modified, resulting in different production possibility sets [11, 23]. In this section, we discuss some other efficiency models in the context of DEA, such as the BCC model, additive model, and a VDEA model.

BCC model

Firstly, let us focus on the BCC DEA model [10]. It assumes the variable return to the scale of activities. In this model, only the convex combinations of the existing DMUs are feasible. Figure 2.3 compares the efficiency frontiers in a CCR and BCC models for the considered example.

As mentioned before, the BCC model allows only the convex combination of units, i.e., the sum of their contributions in the artificial DMU must be equal to one ($\sum_{k=1}^K \lambda_k = 1$).

The input-oriented model to determine the BCC efficiency in the units' combinations perspective (dual formulation) is the following:

$$\begin{aligned}
& \min \theta \\
& \text{subject to: } \sum_{k=1}^K \lambda_k x_{m,k} \leq \theta x_{m,o}, \quad \text{for } m = 1, 2, \dots, M, \\
& \quad \quad \quad \sum_{k=1}^K \lambda_k y_{n,k} \geq y_{n,o}, \quad \text{for } n = 1, 2, \dots, N, \\
& \quad \quad \quad \sum_{k=1}^K \lambda_k = 1, \\
& \quad \quad \quad \lambda_k \geq 0, \quad \text{for } k = 1, 2, \dots, K.
\end{aligned} \tag{2.17}$$

Note that the only difference from the CCR model (see Equation 2.6) is the additional constraint, which ensures that the artificial DMUs is the convex combination of units, in contrast to the conical combination in the CCR model. Similarly to the CCR model, the unit is deemed weakly efficient if the optimal value of the objective function (θ^*) is equal to one. In the second stage, the fully efficient units, in terms of the BCC model, can be determined with the analogous model to the second phase CCR model (Equation 2.8), with additional constraint, which normalizes the sum of variables λ_k . It is worth noticing that all units which are CCR efficient are also efficient in terms of the BCC model, but not the opposite.

The primal formulation of the input-oriented BCC model is the following:

$$\begin{aligned}
& \max \quad E_o = \mu_0 + \sum_{n=1}^N \mu_n y_{n,o} \\
& \text{subject to: } \sum_{m=1}^M \nu_m x_{m,o} = 1, \\
& \quad \quad \quad \mu_0 + \sum_{n=1}^N \mu_n y_{n,k} \leq \sum_{m=1}^M \nu_m x_{m,k}, \quad \text{for } k = 1, 2, \dots, K, \\
& \quad \quad \quad \mu_n, \nu_m \geq 0, \quad \text{for } n = 1, 2, \dots, N, \quad m = 1, 2, \dots, M, \\
& \quad \quad \quad \mu_0 \text{ free.}
\end{aligned} \tag{2.18}$$

In this model, when compared to the CCR model, the additional variable μ_0 is added to the objective function and the normalizing constraints. Note that this variable is free in sign, i.e., it may take both positive or negative values. Again, the DMUs, for which the optimal objective function is equal to one, are weakly efficient. Moreover, if for some DMU_o , in any optimal solution of this model, all weights μ_n, ν_m are non-zero, then DMU_o is strongly (fully) efficient.

Analogously, we can formulate the LP models for output-oriented problems.

Additive model

In the ratio-based efficiency models discussed before, choosing the input or output orientation is necessary. In the additive model [15], both orientations are combined into a

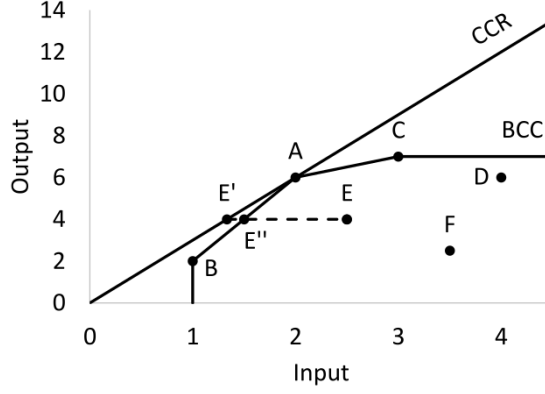


Figure 2.3: Efficiency frontier in CCR and BCC models of Data Envelopment Analysis.

single model. In the further part of this section, we present the additive model for the constant return to scale. However, it can be easily adapted to the problems with variable returns to scale [5]. This model aims to maximize the L_1 distance from the analyzed DMU_o to the efficient frontier. Such a goal is obtained by maximizing the sum of slacks for all inputs and outputs:

$$\begin{aligned}
 & \max \sum_{m=1}^M s_m^+ + \sum_{n=1}^N s_n^- \\
 \text{subject to: } & \sum_{k=1}^K \lambda_k x_{m,k} - s_m^+ = x_{m,o}, & \text{for } m = 1, 2, \dots, M, \\
 & \sum_{k=1}^K \lambda_k y_{n,k} - s_n^- = y_{n,o}, & \text{for } n = 1, 2, \dots, N, \\
 & \lambda_k \geq 0, & \text{for } k = 1, 2, \dots, K, \\
 & s_m^+ \geq 0, s_n^- \geq 0 & \text{for } m = 1, 2, \dots, M, n = 1, 2, \dots, N.
 \end{aligned} \tag{2.19}$$

The DMU_o is efficient, in terms of the additive model, if all slacks (s_m^+ and s_n^-), in the optimal solution, are equal to zero. Moreover, the units deemed efficient with the additive model are the same as those fully efficient in the CCR model [2].

The LP model of the additive model in the efficiency scores (primal) perspective is the following:

$$\begin{aligned}
 & \min \sum_{m=1}^M \nu_m x_{m,o} - \sum_{n=1}^N \mu_n y_{n,o} \\
 \text{subject to: } & \sum_{n=1}^N \mu_n y_{n,k} \leq \sum_{m=1}^M \nu_m x_{m,k}, & \text{for } k = 1, 2, \dots, K, \\
 & \mu_n, \nu_m \geq 0, & \text{for } n = 1, 2, \dots, N, m = 1, 2, \dots, M.
 \end{aligned} \tag{2.20}$$

Example The LP models for evaluating the DMU E from the considered data set, using the additive DEA model are presented below.

Primal formulation:

$$\begin{aligned}
& \min 2.5\nu_{in} - 4\mu_{out} \\
& \text{s.t. } 6\mu_{out} \leq 2\nu_{in}, \\
& \quad 2\mu_{out} \leq \nu_{in}, \\
& \quad 7\mu_{out} \leq 3\nu_{in}, \\
& \quad 6\mu_{out} \leq 4\nu_{in}, \\
& \quad 4\mu_{out} \leq 2.5\nu_{in}, \\
& \quad 2.5\mu_{out} \leq 3.5\nu_{in}.
\end{aligned}$$

Dual formulation:

$$\begin{aligned}
& \max s_{in}^+ + s_{out}^- \\
& \text{s.t. } 2\lambda_A + \lambda_B + 3\lambda_C + 4\lambda_D + 2.5\lambda_E + 3.5\lambda_F + s_{in}^+ = 2.5, \\
& \quad 6\lambda_A + 2\lambda_B + 7\lambda_C + 6\lambda_D + 4\lambda_E + 2.5\lambda_F - s_{out}^- = 4, \\
& \quad \lambda_A, \lambda_B, \lambda_C, \lambda_D, \lambda_E, \lambda_F \geq 0, \\
& \quad s_{in}^+, s_{out}^- \geq 0.
\end{aligned} \tag{2.22}$$

Some other DEA models, based on the additive models, are the weighted additive model [4], a slack-based measure of efficiency [80], and the VDEA model [26]. The latter is described in more detail in the next section.

Value-based additive DEA model

Another efficiency model, called Value-based additive DEA (VDEA) [26, 38], is inspired by the additive DEA model and the MAVT [34, 48] gathered from MCDA. In this model, for each factor q (input and output), we define a value function (u_q) that transforms the original input and output values into the values from the range $[0, 1]$. To maintain the spirit of DEA, the increase of inputs is undesirable, so the value function for input factors must be non-increasing. For outputs, the value function must be non-decreasing, as the increase of outputs in a desirable situation. The examples of the value functions proposed for the considered example data set are presented in Figure 2.4. For example, for DMU A the input value is equal to 2, so its utility is equal to 0.8 ($u_{in}(A) = u_{in}(2) = 0.8$), while its output value is equal to 6, which corresponds to the utility value 0.7 ($u_{out}(A) = u_{out}(6) = 0.7$). The efficiency score of DMU_o , within the VDEA model, is a weighted sum of values assigned to all factors:

$$E_o = \sum_{q=1}^Q w_q u_q(DMU_o), \tag{2.23}$$

where Q is the number of factors ($Q = N + M$). The sum of weights assigned to inputs and outputs must be equal to one. This ensures that the efficiency score is always between 0 for an anty-ideal DMU and 1 for the ideal one.

To evaluate the DMU_o under this model, we compare them to all other units and minimize the maximal distance of DMU_o to any other unit in terms of the efficiency score:

$$\begin{aligned}
& \min d_o \\
& \text{subject to: } \sum_{q=1}^Q w_q u_q(DMU_k) - \sum_{q=1}^Q w_q u_q(DMU_o) \leq d_o, \quad \text{for } k = 1, 2, \dots, K, \\
& \quad \sum_{q=1}^Q w_q = 1, \\
& \quad d_o \geq 0, \\
& \quad w_q \geq 0, \quad \text{for } q = 1, 2, \dots, Q.
\end{aligned} \tag{2.24}$$

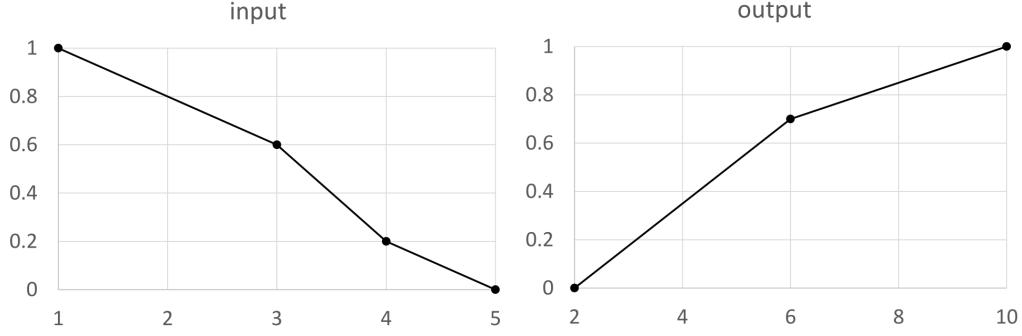


Figure 2.4: The value functions assigned to the input and output in the example data set.

The DMUs, for which the minimal distance (d_o^*) in the optimal solution of this model is equal to 0, are efficient. If $d_o^* > 0$, then DMU_o is inefficient.

Example To evaluate the efficiency of the example DMU E , with the VDEA model, the following model need to be solved:

$$\begin{aligned}
& \min d_E \\
& \text{subject to: } w_{in}u_{in}(2) + w_{out}u_{out}(6) - (w_{in}u_{in}(2.5) + w_{out}u_{out}(4)) \leq d_E, \\
& \quad w_{in}u_{in}(1) + w_{out}u_{out}(2) - (w_{in}u_{in}(2.5) + w_{out}u_{out}(4)) \leq d_E, \\
& \quad w_{in}u_{in}(3) + w_{out}u_{out}(7) - (w_{in}u_{in}(2.5) + w_{out}u_{out}(4)) \leq d_E, \\
& \quad w_{in}u_{in}(4) + w_{out}u_{out}(6) - (w_{in}u_{in}(2.5) + w_{out}u_{out}(4)) \leq d_E, \\
& \quad w_{in}u_{in}(2.5) + w_{out}u_{out}(4) - (w_{in}u_{in}(2.5) + w_{out}u_{out}(4)) \leq d_E, \\
& \quad w_{in}u_{in}(3.5) + w_{out}u_{out}(2.5) - (w_{in}u_{in}(2.5) + w_{out}u_{out}(4)) \leq d_E, \\
& \quad d_E \geq 0, \\
& \quad w_{in}, w_{out} \geq 0.
\end{aligned} \tag{2.25}$$

The minimal distance d_E^* obtained with this model is equal to 0.117, so DMU E is inefficient.

2.3 Super-efficiency

The standard DEA method, described in Section 2.1, allows us to determine the efficiency score of DMUs and divide their set into two subsets: efficient and inefficient ones. Such an approach is not capable of comparing the efficient units. One of the DEA extensions which deal with this problem is super-efficiency [6]. This approach eliminates the considered unit from the data set and measures its distance from the efficient frontier constructed with the remaining DMUs. This idea is presented in Figure 2.5. In this case, the super-efficiency of the DMU A is calculated as follows:

$$SE_o = \frac{|OA'|}{|OA|}. \tag{2.26}$$

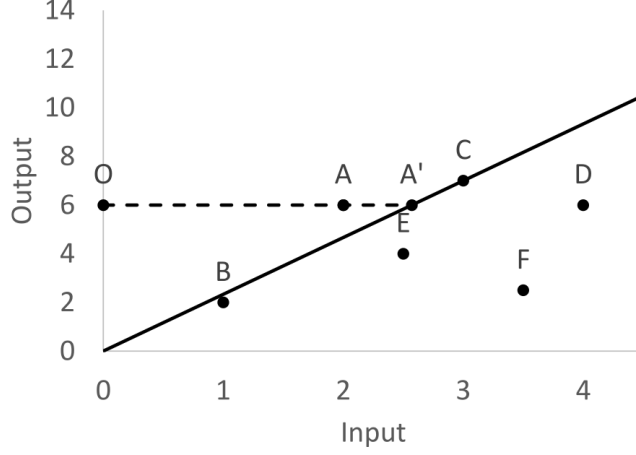


Figure 2.5: Super-efficiency for DMU A from the considered example.

This approach allows for the efficient units to attain a super-efficiency score greater than 1, which eliminates some ties between them. The super-efficiency measure also has some other practical usage. See [66, 85] for further details.

The following LP model is an input-oriented variant for determining the super-efficiency for DMU_o . It is similar to the standard primal input-oriented model. However, the constraint which limits the efficiency score of the evaluated unit is removed, so it can attain the super-efficiency score greater than 1:

$$\begin{aligned}
\max SE_o &= \sum_{n=1}^N \mu_n y_{n,o} \\
\text{s.t.} \quad &\sum_{m=1}^M \nu_m x_{m,o} = 1, \\
&\sum_{n=1}^N \mu_n y_{n,k} \leq \sum_{m=1}^M \nu_m x_{m,k}, && \text{for } k = 1, 2, \dots, K, k \neq o, \\
&\mu_n, \nu_m \geq 0, && \text{for } n = 1, 2, \dots, N, m = 1, 2, \dots, M.
\end{aligned} \tag{2.27}$$

The input-oriented model for super-efficiency calculation from the perspective of the efficiency frontier is presented below:

$$\begin{aligned}
\min \theta \\
\text{subject to:} \quad &\sum_{k=1, k \neq o}^K \lambda_k x_{m,k} \leq \theta x_{m,o}, && \text{for } m = 1, 2, \dots, M, \\
&\sum_{k=1, k \neq o}^K \lambda_k y_{n,k} \geq y_{n,o}, && \text{for } n = 1, 2, \dots, N, \\
&\lambda_k \geq 0, && \text{for } k = 1, 2, \dots, K, k \neq o, \\
&\theta \geq 0.
\end{aligned} \tag{2.28}$$

It is possible to construct analogous models for the output-oriented problem.

Example: To determine the super-efficiency of the efficient DMU A , one the following models must be solved:

Primal formulation:

$$\begin{aligned}
& \max 6\mu_{out} \\
& \text{s.t.}: 2\nu_{in} = 1, \\
& \quad 2\mu_{out} \leq \nu_{in} \quad (B), \\
& \quad 7\mu_{out} \leq 3\nu_{in} \quad (C), \\
& \quad 6\mu_{out} \leq 4\nu_{in} \quad (D), \\
& \quad 4\mu_{out} \leq 2.5\nu_{in} \quad (E), \\
& \quad 2.5\mu_{out} \leq 3.5\nu_{in} \quad (F).
\end{aligned}
\tag{2.29}$$

Dual formulation:

$$\begin{aligned}
& \min \theta \\
& \text{s.t.} \lambda_B + 3\lambda_C + 4\lambda_D + 2.5\lambda_E + 3.5\lambda_F \leq 2\theta, \\
& \quad 2\lambda_B + 7\lambda_C + 6\lambda_D + 4\lambda_E + 2.5\lambda_F \geq 6, \quad (2.30) \\
& \quad \lambda_B, \lambda_C, \lambda_D, \lambda_E, \lambda_F, \\
& \quad \theta \geq 0.
\end{aligned}$$

The idea of the super-efficiency was also developed for other efficiency models, e.g., the VDEA [39], the slacks-based efficiency model [81], and the additive model [30].

2.4 Cross-efficiency

Another measure, which deals with the incomparability of the efficient units, is a cross-efficiency [72]. The main idea of this procedure is to evaluate the considered DMU_o with the most favorable weight vectors obtained for other units from the data set.

In the first step of calculating the cross-efficiency of any DMU, we need to determine the most favorable weight vectors for each unit. It can be done by solving the primal formulation of the standard model of DEA. Depending on the chosen model and orientation, it is one of the models presented in Equations 2.4, 2.13, 2.17 or similar. The set of the optimal weights for the DMU_k is denoted as $(\nu_k^*, \mu_k^*), \nu_k^* = \{\nu_{1,k}^*, \nu_{2,k}^*, \dots, \nu_{M,k}^*\}, \mu_k^* = \{\mu_{1,k}^*, \mu_{2,k}^*, \dots, \mu_{N,k}^*\}$. The cross-efficiency score of DMU_o under the optimal weights of DMU_k is computed as follows:

$$CE_{o,k} = \frac{\sum_{n=1}^N \mu_{n,k}^* y_{n,o}}{\sum_{m=1}^M \nu_{m,k}^* x_{m,o}}. \tag{2.31}$$

Having determined the cross-efficiency matrix $(CE_{i,j}, i = 1, 2, \dots, K, j = 1, 2, \dots, K)$, the DMUs can be compared by the arithmetical mean of their cross-efficiency scores (\overline{CE}_o) , i.e.

$$\overline{CE}_o = \frac{\sum_{k=1}^K CE_{o,k}}{K}. \tag{2.32}$$

The aggregation of the individual cross-efficiency scores has three main drawbacks [88]. First, the set of weights for which the efficiency score is the best may not be unique, which implies some randomness in the average cross-efficiency result [29]. Second, the correlation between the weights and the cross-efficiency score is lost in the averaging procedure, so the decision-makers can be deprived of valuable information about the possible improvements [87]. Third, the means of the cross-efficiency scores are not Pareto optimal [86]. The further development of the cross-efficiency measures resulted in multiple extensions which deal with these problems [7, 52, 87].

2.5 Robustness analysis in Data Envelopment Analysis

In this section, we discuss the existing approaches of robustness analysis in the context of DEA. The idea of the robustness analysis has its origins in the MCDA and accounts for the uncertainty in real-world problems [46]. The outcomes of some methods are considered robust if they are true for the majority of the feasible combinations of parameters [67, 83]. The results of the robustness analysis are often helpful to the decision makers and guide them to enrich the provided preference information and narrow down the space of the feasible solutions to obtain results more robust.

In the context of DEA, the robustness analysis approaches are focused on different aspects of the analyzed data set, e.g., the changes of the input and output values, adding or removing factors, or choosing the efficiency model [89, 18]. In this dissertation, we propose the robustness analysis, which explores the feasible space of the weight vectors. We build on the two concepts.

On the one hand, [70] proposed the Ratio-based Efficiency Analysis (REA), which explores the whole set of the input/output weight vectors and provides results in three perspectives: the efficiency bounds, i.e., the greatest and the lowest possible efficiencies of a unit compared to a subset of other DMUs, the dominance relations between pairs of DMUs and the ranking intervals (the range of efficiency ranks attained by a DMU). Unlike the standard DEA, the outcomes are derived from pairwise comparisons rather than measuring their distance to the efficient frontier.

On the other hand, [59] introduces the stochastic approach, based on the Stochastic Multicriteria Acceptability Analysis (SMAA) [58, 64], called Stochastic Multicriteria Acceptability Analysis for Data Envelopment Analysis (SMAA-D). The proposed method handles the uncertainties of the input/output weights and performances. It provides the stochastic index, which calculates how often the given DMU attains the specific position in the efficiency ranking.

Chapter 3

Robustness analysis framework for Data Envelopment Analysis

In this chapter, we describe the robustness analysis framework for DEA proposed in this dissertation.

3.1 Basic concepts

In Data Envelopment Analysis, the standard approach allows us to find the efficiency score for each DMU considering its most favorable input-output weight vector. Such efficiency score allows dividing the set of analyzed DMUs into efficient and inefficient ones. However, the number of efficient units can be relatively high, especially when the number of indicators (inputs and output) increases. We cannot compare the efficient units as they all attain the same efficiency score equal to 1. Moreover, even for the inefficient units, their comparison may be irrelevant because of the different weight vectors used to assess the performance of each DMU. Publication P1 presents the novel framework for robustness analysis for DEA with a ratio-based efficiency model. The proposed methods concern three points of view: efficiency scores, pairwise efficiency preference relations, and efficiency ranks.

The efficiency evaluation using the proposed framework is conducted in two complementary ways. Firstly, we use the LP techniques to determine the exact outcomes in each point of view: extreme efficiency scores, extreme efficiency ranks, and to verify the truth of the necessary and possible efficiency preference relations. Finally, the Monte Carlo simulation is used to find the estimated stochastic indices based on the sample of the feasible weight vector space.

The combination of both types of results is beneficial for further analysis. On the one hand, mathematical programming methods provide information on what happens in the most and the least favorable scenario for a given unit. However, the difference between the extreme efficiencies or ranks can be, in many cases, large. Moreover, these extreme values occur only for a single, particular weight vector and may be very far from the efficiencies and ranks attained in the average situation. Similarly, the possible preference relation may be true for the majority of pairs, making many DMUs incomparable in terms of a robust preference. In this situation, the stochastic analysis can give some additional information about the distributions of the efficiency scores and ranks, the

estimated average score and rank, and the probability of the preference of one DMU over the other. On the other hand, the stochastic indices can be estimated with high accuracy. However, they are not exact. In particular, the chance of hitting the weight vector, for which a unit attains the extreme score, is very low. Analogously, the pairwise efficiency outranking index for some pair can be equal to one, which does not confirm that one DMU is always preferred to another. Thus, the confrontation of the stochastic indices with the exact outcomes is desirable.

To determine the stochastic indices for problems with ratio-based efficiency model, the following algorithm is applied. As the weight vector space is unbounded, we added the constraints which normalize the weights in the following form:

$$\sum_{n=1}^N \mu_n = \sum_{m=1}^M \nu_m = 1. \quad (3.1)$$

Then, we sample a set of weight vectors from this space using the Hit-And-Run algorithm [78]. Following the SMAA-D [59], we use the uniform distribution for sampling. However, it can generally be replaced with any probability distribution with a joint density function in the feasible weight space.

After obtaining the input/output weight vectors, the efficiency scores for each DMU are computed. After that, we normalize them by dividing them by the maximal obtained efficiency, which transforms the efficiency scores into the interval between zero and one, as in the traditional DEA approach. Thus obtained efficiency scores are analyzed, providing the estimates of the distributions of the efficiency score and efficiency ranks over the feasible weight vector space. Moreover, we estimate the share of weight vectors for which some DMU is preferred to another.

3.2 Efficiency scores

In this section, we describe the robustness measures considering the viewpoint of the efficiency scores. To obtain the extreme (minimal and maximal) efficiency scores, the LP models were proposed. The maximal efficiency score can be determined by solving the original Charnes, Cooper and Rhodes (CCR) DEA model (see Equation 2.4). To determine the worst efficiency score for DMU_o ($E_{o,*}$) the following LP model must be solved:

$$\begin{aligned} \min E_{o,*} &= \sum_{n=1}^N \mu_n y_{n,o} \\ \text{subject to: } & \sum_{m=1}^M \nu_m x_{m,o} = 1, \\ & \sum_{n=1}^N \mu_n y_{n,k} \geq \sum_{m=1}^M \nu_m x_{m,k} - C(1 - b_k), \quad \text{for } k = 1, 2, \dots, K, \\ & \sum_{k=1}^K b_k \geq 1, \\ & b_k \in \{0, 1\}, \quad \text{for } k = 1, 2, \dots, K, \\ & (\mu, \nu) \in S_w, \end{aligned} \quad (3.2)$$

where C is a large positive constant. In this case, we look for the least advantageous weight vector for DMU_o . Its efficiency score is minimized with the restriction that at least one DMU remains efficient ($E \geq 1$). This condition is ensured by using the binary variables b_k . If b_k for some DMU_k is set to 1 then the component $C(1 - b_k)$ is equal to 0 and the constraint takes form $\sum_{n=1}^N \mu_n y_{n,k} \geq \sum_{m=1}^M \nu_m x_{m,k}$. Otherwise (is $b_k = 0$), $C(1 - b_k) = C$, and the right side of the constraint is small enough to make it always met. By adding the constraint $\sum_{k=1}^K b_k \geq 1$, we ensure that for at least one DMU_k , the constraint holds and the efficiency score is not worse than 1.

To determine the distribution of the efficiency scores over the weight vector space, we estimate the EAIIs for all DMUs. The Efficiency Acceptability Interval Index for a DMU_o and an interval b_i ($EAI(DMU_o, b_i)$) is the share of the feasible weight vectors $(\mu, \nu) \in S_w$, for which the DMU_o attains the efficiency score in the interval b_i . The intervals (buckets) $b_i, i = 1, 2, \dots, B$, are defined by their extreme values $b_{i,*}$ and b_i^* , i.e., $b_i = (b_{i,*}, b_i^*]$ with the proviso that b_1 is also left-closed. They must be disjoint and cover the whole efficiency space, i.e., $b_i \cap b_j = \emptyset$, if $i \neq j$ and $\bigcup_{i=1}^B b_i = [0, 1]$. Moreover, by default, we assume that the buckets have equal widths. However, it is possible to construct buckets with different widths. For each DMU_o , the sum of its EAIIs is equal to one, i.e.:

$$\sum_{i=1}^B EAI(DMU_o, b_i) = 1. \quad (3.3)$$

Moreover, the stochastic analysis is enriched by determining some additional measures for each DMU_o , such as the extreme efficiencies (E_o^* and $E'_{o,*}$) observed with Monte Carlo simulation and the estimated expected efficiency score EE'_o , defined as:

$$EE'_o = \frac{\sum_{(\mu, \nu) \in S_w^S} E_o(\mu, \nu)}{W}, \quad (3.4)$$

where S_w^S is the set weight vector samples and W is the number of samples.

3.3 Efficiency ranks

For a given DMU_o and the weight vector $(\mu, \nu) \in S_w$, the efficiency rank of DMU_o (R_o) is defined as the number of DMUs for which the efficiency score with this weight vector is greater than the efficiency score of DMU_o increased by one, i.e.:

$$R_o = 1 + \sum_{k=1, k \neq o}^K h(o, k, (\mu, \nu)), \text{ where} \quad (3.5)$$

$$h(o, k, (\mu, \nu)) = \begin{cases} 1, & \text{if } E_k(\mu, \nu) > E_o(\mu, \nu) \\ 0, & \text{otherwise.} \end{cases} \quad (3.6)$$

To indicate the best (minimal) efficiency rank (R_o^*) for DMU_o , the following LP model

need to be solved:

$$\begin{aligned}
& \min R_o^* = 1 + \sum_{k=1, k \neq o}^K b_k \\
& \text{subject to: } \sum_{m=1}^M \nu_m x_{m,o} = 1, \\
& \sum_{n=1}^N \mu_n y_{n,o} = 1, \\
& \sum_{n=1}^N \mu_n y_{n,k} \leq \sum_{m=1}^M \nu_m x_{m,k} + C b_k, \quad \text{for } k = 1, 2, \dots, K, k \neq o \\
& b_k \in \{0, 1\} \quad \text{for } k = 1, 2, \dots, K, k \neq o \\
& (\mu, \nu) \in S_w.
\end{aligned} \tag{3.7}$$

In this case, we fix the efficiency score of DMU_o to the value of 1 and minimize the number of DMUs for which the efficiency score is greater than 1. If, for some weight vector, the efficiency score cannot be less than or equal to 1 ($\sum_{n=1}^N \mu_n y_{n,k} \leq \sum_{m=1}^M \nu_m x_{m,k}$), then the binary variable b_k is set to 1. By multiplying it by a large constant C , it is ensured that the constraint is satisfied. The variables b_k set to 1 in the optimal solution of the model identify the DMUs better than DMU_o . If $E_k(\mu, \nu) \leq E_o(\mu, \nu)$, then b_k is instantiated with zero. By minimizing the sum of the variables b_k , the solver assigns as many zeros as possible. This sum, increased by one, gives the minimal (best) efficiency rank of DMU_o .

The worst efficiency rank for DMU_o ($R_{o,*}$) can be determined with the following LP model:

$$\begin{aligned}
& \max R_{o,*} = 1 + \sum_{k=1, k \neq o}^K b_k \\
& \text{subject to: } \sum_{m=1}^M \nu_m x_{m,o} = 1, \\
& \sum_{n=1}^N \mu_n y_{n,o} = 1, \\
& \sum_{m=1}^M \nu_m x_{m,k} \leq \sum_{n=1}^N \mu_n y_{n,k} + C(1 - b_k), \quad \text{for } k = 1, 2, \dots, K, k \neq o \\
& b_k \in \{0, 1\} \quad \text{for } k = 1, 2, \dots, K, k \neq o \\
& (\mu, \nu) \in S_w.
\end{aligned} \tag{3.8}$$

Again, we assign $E_o(\mu, \nu) = 1$, but in this case, we search for the maximal number of DMUs (DMU_k), for which $E_k(\mu, \nu) \geq E_o(\mu, \nu)$. If $E_k(\mu, \nu)$ is not worse than $E_o(\mu, \nu)$, then a binary variable b_k is instantiated with one. Otherwise, if the efficiency score of DMU_k cannot be better than or equal to the efficiency score of DMU_o , then b_k is instantiated with zero. The sum of the binary variables b_k , increased by one, gives the worst possible rank for DMU_o .

To enrich the analysis from the viewpoint of efficiency ranks, we estimate the Efficiency Rank Acceptability Indices (ERAI) for all DMUs. ERAI for a given DMU_o and

a rank r ($ERAI(DMU_o, r)$) is defined as the share of feasible weight vectors $(\mu, \nu) \in S_w$, for which the DMU_o attains r th position in the efficiency ranking. For each DMU ERAIs have the following property:

$$\sum_{k=1}^K ERAI(DMU_o, k) = 1. \quad (3.9)$$

Moreover, we also use the Monte Carlo simulation to estimate the expected efficiency rank (ER'_o) for each DMU_o :

$$ER'_o = \sum_{k=1}^K k \cdot ERAI(DMU_o, k). \quad (3.10)$$

3.4 Preference relations

From the viewpoint of pairwise efficiency comparisons, we define two efficiency preference relations:

- possible efficiency preference relation, (\succsim_E^P), for a pair of DMUs (DMU_o, DMU_k), is verified if there exists at least one weight vector, for which the efficiency score of DMU_o is not worse than the efficiency score of DMU_k ,
- necessary efficiency preference relation, (\succsim_E^N), for a pair of DMUs (DMU_o, DMU_k), is verified if, for all feasible weight vectors, the efficiency score of DMU_o is better than or equal to the efficiency score of DMU_k .

To verify if the possible efficiency preference relation is held for a pair of DMUs (DMU_o, DMU_k), the following model must be solved:

$$\begin{aligned} \max E_o &= \sum_{n=1}^N \mu_n y_{n,o} \\ \text{subject to: } & \sum_{m=1}^M \nu_m x_{m,o} = 1, \\ & \sum_{n=1}^N \mu_n y_{n,k} = \sum_{m=1}^M \nu_m x_{m,k}, \\ & (\mu, \nu) \in S_w. \end{aligned} \quad (3.11)$$

In this model, we maximize the efficiency score of DMU_o with the restriction that $E_k = 1$. If the objective value of the optimal solution of this problem (E_o^{max}) is not worse than 1, then exists some weight vector (μ, ν) for which $E_o(\mu, \nu) \geq E_k(\mu, \nu)$, so $E_o \succsim_E^P E_k$.

The same LP model, but with the opposite optimization direction, allows us to assess if the necessary efficiency preference holds for a pair (DMU_o, DMU_k). In this case, if the minimal objective value is greater than or equal to 1, then $E_o(\mu, \nu) \geq E_k(\mu, \nu)$ for all $(\mu, \nu) \in S_w$, i.e., $E_o \succsim_E^N E_k$.

For a pair of DMUs, (DMU_o, DMU_k), we define the Pairwise Efficiency Outranking Index $PEOI(DMU_o, DMU_k)$ as the share of feasible weight vectors for which the efficiency score of DMU_o is not worse than the efficiency score of DMU_k . The PEOIs for

all DMUs are estimated with Monte Carlo simulation. We can point out the following properties of the PEOIs:

- for each $DMU_o \in \mathcal{D}$, $PEOI(DMU_o, DMU_o) = 1$, i.e., any unit is always as good as itself;
- For a pair $(DMU_o, DMU_k) \in \mathcal{D} \times \mathcal{D}$,

$$1 \leq PEOI(DMU_o, DMU_k) + PEOI(DMU_k, DMU_o) \leq 2,$$

i.e., the sum of PEOIs for a pair of units and the same pair reversed is always within the interval $[1, 2]$, which results from the possible ties in efficiency scores. If there are no ties, such a sum is equal to 1. On contrary, if the efficiency scores of both DMUs is always the same, then $PEOI(DMU_o, DMU_k) + PEOI(DMU_k, DMU_o) = 2$.

3.5 Robustness analysis for Value-based additive efficiency model

In publication P4 the robustness analysis framework for the VDEA model was proposed. Similarly to the ratio-based efficiency model, in terms of VDEA, we perform the analysis using the mathematical programming (extreme outcomes) and Monte Carlo simulation (stochastic distributions), and again, we provide outcomes on three different points of view: efficiency scores, efficiency ranks and DMUs' pairwise preferences.

In the context of VDEA, in the perspective of the efficiency scores, we may consider the relative distances to the best DMU and the absolute values. When it comes to the efficiency distance, we search for the interval $[d_{o,*}, d_o^*]$ delimited by the minimal (the best) $d_{o,*}$ and the maximal (the worst) d_o^* distance to the best DMU. The best distance is computed using the standard formulation of VDEA, described in Section 2.2. The maximal distance can be determined by solving the following LP model:

$$\begin{aligned} & \max d_o \\ \text{subject to: } & \sum_{q=1}^Q w_q u_q(DMU_k) - d_o \geq \sum_{q=1}^Q w_q u_q(DMU_o) - C(1 - b_k), \text{ for } k = 1, \dots, K, \\ & d_o \geq 0, \\ & \sum_{k=1, k \neq o}^K b_k = 1, \\ & b_k \in \{0, 1\}, \text{ for } k = 1, 2, \dots, K; k \neq o, \\ & \left. \begin{aligned} & \sum_{q=1}^Q w_q = 1, \\ & w_q \geq 0, \text{ for } q = 1, \dots, Q, \\ & \mathbf{w} \in S_w. \end{aligned} \right\} \mathcal{W} \end{aligned}$$

In this model, we maximize the distance of DMU_o from some other DMU. The first four constraints ensure that d_o equals the difference between DMU_o and some DMU_k ($o \neq k$). When b_k is equal to 0 then the first constraint is always satisfied, while for $b_k = 1$ the component $C(1 - b_k) = 0$ and $d_o = E_k - E_o$. It is required that $b_k = 1$ for some $DMU_k, k \neq o$.

When considering the absolute values of efficiencies, one wants to determine the extreme, i.e., minimal and maximal possible efficiency scores for analyzed DMUs. Such extreme efficiency scores for a DMU_o can be determined by solving the following model:

$$\min / \max \sum_{q=1}^Q w_q u_q(DMU_o), \text{ s.t. } \mathcal{W}.$$

This time we optimize the efficiency score of DMU_o subject to the constraints defining a set of admissible inputs and output weights.

When considering the other perspectives of the exact analysis, we construct mathematical programming models which allow us to determine the extreme efficiency ranks and to verify the truth of the necessary and possible efficiency preference relations for pairs of DMUs. The proposed models are analogous to those proposed for the ratio-based efficiency model. For brevity, they are not presented in this dissertation.

The stochastic robustness analysis for VDEA is also performed similarly to these proposed for the ratio-based model, which is described in Section 3.1. Again, we use the Hit-And-Run algorithm to obtain a representative set of weight vectors. However, for the VDEA model, we ensure that the sum of the weights is equal to one for all indicators, i.e.

$$\sum_{q=1}^Q w_q = 1. \quad (3.12)$$

After generating the weight samples, the efficiency scores for all DMUs are computed using the standard equation: $E_o = \sum_{q=1}^Q w_q u_q(DMU_o)$. Note that the application of the marginal value functions and the restriction on the sum of the weights already make the efficiency score lay within the range $[0 - 1]$, so no normalization is needed.

Based on the efficiency scores, obtained for weight samples, we compute the same stochastic acceptability indices and measures as for the ratio-based model, i.e., EAII, ERAI, PEOI, expected efficiency score and expected efficiency rank. Moreover, we also determine the Distance Acceptability Interval Indices (DAIIs) for all DMUs. Analogously to EAII, the DAII for a given DMU_o and a bucket b_i is defined as the share of the feasible weight vectors $(\mu, \nu) \in S_w$ for which the distance of DMU_o to the best DMU is within the interval b_i . The construction of those intervals is the same as for EAII. Moreover, based on the obtained weight vector samples, for a given DMU_o , we also estimate the expected value of the distance to the best unit (Ed_o).

3.6 Interdependencies between the robust and stochastic results

The extreme, necessary, and possible results from the mathematical programming influence the stochastic indices in the following way. For each $DMU_o \in \mathcal{D}$:

- if $b_{i,*} > E_o^*$ or $b_i^* < E_{o,*}$, then $EAII(DMU_o, b_i) = 0$, i.e., if the whole subinterval b_i is outside the range of extreme efficiency scores, then the EAII must be equal to zero;
- $\sum_{i: b_{i,*} \leq E_o^* \wedge b_i^* \geq E_{o,*}} EAII(DMU_o, b_i) = 1$, i.e., the sum of the EAII corresponding to the intervals with non-empty intersection with the $[E_o^*, E_{o,*}]$ is equal to one;

- $[E'_{o,*}, E'^*_{o,*}] \subseteq [E_{o,*}, E^*_{o,*}]$, i.e., the interval defined by the estimated extreme efficiency scores is always a subinterval of the real extreme efficiency scores;
- if $b_{i,*} > d^*_{o,*}$ or $b_i^* < d_{o,*}$, then $DAII(DMU_o, b_i) = 0$, i.e., for subintervals b_i with empty intersection with the extreme distances, the DAII must be equal to zero;
- $\sum_{i: b_{i,*} \leq d^*_{o,*} \wedge b_i^* \geq d_{o,*}} DAII(DMU_o, b_i) = 1$, i.e., the sum of the DAIs corresponding to the intervals with a non-empty intersection with the extreme distance to the best DMU is always equal to one;
- If $DMU_o \succsim_E^N DMU_k$, then $PEOI(DMU_o, DMU_k) = 1$, i.e., for pairs of units, for which the necessary efficiency preference is valid, it has to be confirmed by all feasible weight vectors, thus the PEOI is equal to one;
- $\neg(DMU_o \succsim_E^P DMU_k) \implies PEOI(DMU_o, DMU_k) = 0$, i.e., if the preference for a pair of (DMU_o, DMU_k) is not possible, then the preference may not be confirmed by any of the sampled weight vectors, so $PEOI(DMU_o, DMU_k) = 0$;
- $r : \{r < R_o^* \vee r > R_{o,*}\} \implies ERAI(DMU_o, r) = 0$, i.e., if a rank is outside the extreme ranks range, then the ERAI is equal to zero;
- $\sum_{r=R_o^*}^{R_{o,*}} ERAI(DMU_o, r) = 1$, i.e. the sum of ERAIs for all ranks possible attained by DMU_o must be equal to one.

The inverse relations may not be true because of the stochastic nature of the indices. For example, the estimated $PEOI(DMU_o, DMU_k) = 1$ for some pair of DMUs does not imply that the preference between these units is necessary. It means only that for all sampled weight vectors, DMU_o was not worse than DMU_k . However, this may not be true for some vectors which were not hit in the sampling procedure. The following properties are the only ones that can be indicated between the estimated stochastic indices and the exact outcomes:

- $EAI(DMU_o, b_i) > 0 \implies b_{i,*} \leq E_o^* \wedge b_i^* \geq E_{o,*}$, i.e., if the estimated EAI for some bucket is non-zero, then this bucket must have a non-empty intersection with the extreme efficiency scores;
- $PEOI(DMU_o, DMU_k) > 0 \implies DMU_o \succsim_E^P DMU_k$, i.e., if the PEOI for a pair of units is positive, then the possible preference for this pair is always possible;
- $PEOI(DMU_o, DMU_k) < 1 \implies \neg(DMU_o \succsim_E^N DMU_k)$, i.e., if PEOI is lesser than one, then for some weight vector DMU_o is worse than DMU_k , thus the necessary preference is not true;
- $ERAI(DMU_o, r) > 0 \implies r \geq R_o^* \wedge r \leq R_{o,*}$, i.e., if the ERAI for some rank is non-zero, then this rank must be within the extreme ranks interval.

3.7 Evolution of robust results with incremental specification of weight constraints

In this section, we describe how the robust outcomes change when incrementally introducing the weight constraints. Let us consider a Decision-Maker (DM), who introduces

the weight constraints in s iterations. They denote the nested sets of weight constraints provided by the DM as $A^1 \subseteq A^2 \subseteq \dots \subseteq A^s$. These constraints translate into the sets of the feasible weight vectors. For the set of constraints $A^t, t = 1, 2, \dots, s$, the set of feasible weight vectors is denoted as S_w^t . These are incrementally constrained, i.e., $S_w^1 \supseteq S_w^2 \supseteq \dots \supseteq S_w^s$. The outcomes obtained in t -th iteration for DMU_o are the following:

- the extreme efficiency scores – $E_o^{*,t}$ and $E_{o,*}^t$;
- the extreme distances to the best unit – $d_{o,*}^t$ and $d_o^{*,t}$;
- the extreme efficiency ranks – $R_o^{*,t}$ and $R_{o,*}^t$;
- the pairwise efficiency preference relations – $\succsim_E^{N,t}$ and $\succsim_E^{P,t}$.

The evolution of the results with an increase of weight constraints is the following:

- $E_o^{*,t} \leq E_o^{*,t-1}$ and $E_{o,*}^t \geq E_{o,*}^{t-1}$, i.e., the extreme efficiency scores in the subsequent iterations are becoming narrower;
- $d_o^{*,t} \leq d_o^{*,t-1}$ and $d_{o,*}^t \geq d_{o,*}^{t-1}$, i.e., the extreme distances to the best unit may become narrower when introducing new constraints;
- $R_o^{*,t} \geq R_o^{*,t-1}$ and $R_{o,*}^t \leq R_{o,*}^{t-1}$, i.e., the ranking intervals obtained in subsequent iterations are nested;
- $\succsim_E^{N,t} \supseteq \succsim_E^{N,t-1}$ and $\succsim_E^{P,t} \subseteq \succsim_E^{P,t-1}$, i.e., the necessary preference relations may be enriched, while the possible relations may be impoverished.

3.8 Robustness analysis for Data Envelopment Analysis with imprecise information

In Publications P5 and P6, we present the robustness analysis framework for DEA with imprecise information [22, 28, 90]. Publication P5 focuses on the ratio-based model, while in Publication P6, we consider the VDEA model. For both efficiency models, we account for two types of uncertainty. The performance can be defined in the form of interval [39] or in the form of ranking defining only the order of DMUs for a given indicator. Moreover, for VDEA, we also consider the uncertainty at the level of the marginal value functions. Instead of the precise marginal functions, the user can provide the range of the admissible marginal values for cardinal factors. In the further part of this section, we describe how to deal with imprecise information in both mathematical programming and stochastic methods.

Mathematical programming methods

This section presents how the different types of uncertainty are treated in mathematical programming models.

Interval inputs and outputs. Let us consider an interval input in_m and the interval output out_n . We denote the interval of the possible values of in_m of DMU_k as $[x_{m,k,*}, x_{m,k}^*]$. The interval of possible values of out_n for DMU_k is marked as $[y_{n,k,*}, y_{n,k}^*]$. For mathematical programming models, the interval values are replaced with the precise ones representing the optimistic (the most favorable) or pessimistic (the least favorable) scenario for the analyzed unit DMU_o , depending on the considered result type. When identifying the best possible outcome for DMU_o ($SCE = OPT$), the precise input/output values contain the most favorable ones for DMU_o and the least favorable for the remaining DMUs, i.e.:

$$x_{n,k}^{OPT,o} = \begin{cases} x_{m,k,*}, & \text{if } k = o, \\ x_{m,k}^*, & \text{otherwise,} \end{cases} \quad (3.13)$$

$$y_{n,k}^{OPT,o} = \begin{cases} y_{n,k}^*, & \text{if } k = o, \\ y_{n,k,*}, & \text{otherwise.} \end{cases} \quad (3.14)$$

. Analogously, when searching for the worst possible result for DMU_o ($SCE = PES$), the precise inputs and outputs must represent the least favorable scenario for the analyzed unit, i.e., the maximal inputs and the minimal outputs and the most favorable for others (minimal inputs and maximal outputs):

$$x_{m,k}^{PES,o} = \begin{cases} x_{m,k}^*, & \text{if } k = o, \\ x_{m,k,*}, & \text{otherwise,} \end{cases} \quad (3.15)$$

$$y_{n,k}^{PES,o} = \begin{cases} y_{n,k,*}, & \text{if } k = o, \\ y_{n,k}^*, & \text{otherwise.} \end{cases} \quad (3.16)$$

Ordinal factors. For ordinal factors, we ensure that the order defined by performances is preserved. For this purpose, let us introduce the symbol π_q , which denotes the permutation of DMUs that reorders them according to the non-decreasing order of their performances on factor q . For example $\pi_q(1)$ denotes the DMU with the lowest performance on factor q and $x_{q,\pi_q(1)}$ is the value of q th input for this unit. Having defined such permutation, we need to model the monotonicity of this permutation in mathematical programming. In the ratio-based model, we replace the product $\nu_m \cdot x_{m,o}$ or $\mu_n \cdot y_{n,o}$ by a single variable, respectively $X_{m,o}$ or $Y_{n,o}$, which allows omitting the non-linearity of the model. After that, we ensure the monotonous order of these variables according to the permutation π_q , i.e., if $x_{q,\pi_q(k)} = x_{q,\pi_q(k+1)}$, then $X_{q,\pi_q(k)}$ must be equal to $X_{q,\pi_q(k+1)}$, while if $x_{q,\pi_q(k)} < x_{q,\pi_q(k+1)}$, then $\alpha X_{q,\pi_q(k)} \leq X_{q,\pi_q(k+1)}$, for some $\alpha > 1$. The constraints are constructed analogously to output factors. Note that the multiplicative form of the monotonicity constraints maintains the spirit of DEA [90].

For VDEA model, the ordinal inputs and outputs are treated in a similar way. However, the replacement variable does not model the product of the weight and the performance, but the product of the weight and the marginal value for a given factor and unit, i.e., $U_{q,o} = w_q \cdot u_q(DMU_o)$. As the direction of the monotonicity of the marginal value functions is different for inputs and outputs, the constraints must represent the non-descending order for outputs and non-increasing for inputs. Obviously, if $x_{q,\pi_q(k)} = x_{q,\pi_q(k+1)}$ or $y_{q,\pi_q(k)} = y_{q,\pi_q(k+1)}$, then $U_{q,\pi_q(k)} = U_{q,\pi_q(k+1)}$. Otherwise, if q

is an input and $x_{q,\pi_q(k)} < x_{q,\pi_q(k+1)}$, then $U_{q,\pi_q(k)} \geq \alpha U_{q,\pi_q(k+1)}$. If q is an output and $y_{q,\pi_q(k)} < y_{q,\pi_q(k+1)}$, then $\alpha U_{q,\pi_q(k)} \leq U_{q,\pi_q(k+1)}$. Moreover, we ensure that the values assigned to the worst performances ($U_{q,\pi_q(1)}$ for outputs and $U_{q,\pi_q(K)}$ for inputs) are positive (greater than the small value ϵ). The values corresponding to the best performances ($U_{q,\pi_q(K)}$ for outputs and $U_{q,\pi_q(1)}$ for inputs) must not be greater than the weight assigned to the factor q .

Admissible marginal function range. For the value-based efficiency model, we consider the third form of uncertainty: the range of the admissible marginal value functions. Such range is defined by the shapes of two functions that limit the range from the top (u_q^*) and bottom ($u_{q,*}$). The examples of value function ranges, defined for indicators from the case study considered in P6, are presented in Figure 3.1. To deal with such factors in mathematical programming, we introduce the replacement variables $U_{q,o}$ as for ordinal factors. If the considered factor q is also the interval one, the interval performances of this factor must be replaced by the precise ones (as described before). After that, we need to ensure that the values for individual DMUs are between the defined lower and upper bound, i.e., $u_{q,*}(DMU_k) \leq u_q(DMU_k) \leq u_q^*(DMU_k)$. With the replacement variables, such constraints take form: $w_q u_{q,*}(DMU_k) \leq U_{q,k} \leq w_q u_q^*(DMU_k)$. Moreover, we impose monotonicity constraints similar to these introduced for the ordinal factors.

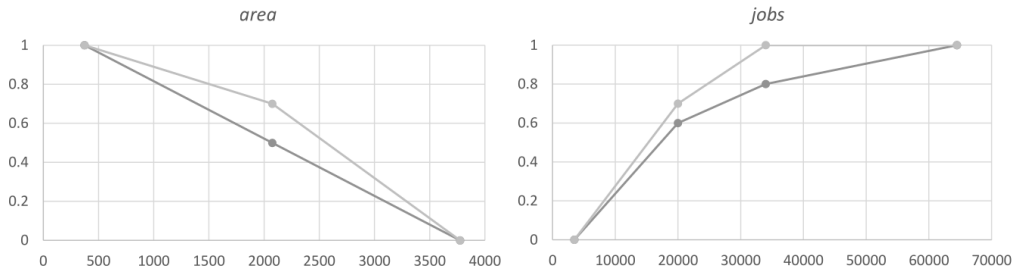


Figure 3.1: Examples of admissible function ranges.

The LP model, which allows determining the minimal distance of DMU_o to the best

The interpretation of constraints in this model is similar to the one described for the minimal distance model. However, this time, we need to apply the permutation of the precise values from the pessimistic scenario for DMU_o ($\pi_{q,o}^{PES}$).

The mathematical models for the remaining robustness analysis perspectives concerning both the VDEA as well as the ratio-based model are constructed analogously and presented in Publications P5 and P6. We omit them in this dissertation to keep it concise.

Stochastic methods

To determine the stochastic indices for DEA with imprecise inputs and outputs, we perform the sampling procedure in multiple stages. Firstly, we run the Hit-And-Run algorithm to obtain the predefined number of weight vector samples in the same way as for precise DEA (see Section 3.1). The next steps depend on the indicator type and efficiency model. For both efficiency models, if the interval indicator is considered, we need to generate the performance samples from the interval $[x_{q,o,*}, x_{q,o}^*]$. For ordinal indicators, we use the SMAA-O approach [59]. We assume that the performances for ordinal factors are drawn from the interval $[0, 1]$ without losing generality. We choose, at random, K values from this range. Then, these values are sorted and treated as a single sample of precise performances of DMUs consistent with the order π_q . When the value-based efficiency model with an admissible function range is applied, in the last step, we generate the marginal value samples. Having precise performances, we randomly choose the marginal values from the interval between the lower and upper functions. In addition, we ensure that the value function generated by each sample is monotonous. For example, if the obtained in previous step precise performance for some DMU_o and factor q is marked as $x_{q,o}^{(t)}$, the marginal value $u_q^{(t)}(DMU_o)$ must satisfy the following constraint: $u_{q,*}(x_{q,o}^{(t)}) \leq u_q^{(t)}(x_{q,o}^{(t)}) \leq u_q^*(x_{q,o}^{(t)})$. Having obtained the weight, performance, and marginal values samples, the efficiency score for each sample is computed according to the chosen efficiency model. Finally, the stochastic indices are computed similarly to standard (precise) DEA.

3.9 Robustness analysis for Data Envelopment Analysis with a hierarchical structure of inputs and outputs

In publication P7, we consider the efficiency of DMUs in the situation, where the indicators (inputs and outputs) are organized in a multiple-level hierarchical structure using the value-based efficiency model.

The problem formulation is the following. Similarly to the standard DEA problems, the DMUs are evaluated using a set of inputs (\mathcal{IN}) and a set of outputs (\mathcal{OUT}). All factors (inputs and outputs) form a set of indicators, denoted as $\mathcal{F} = \{f_1, f_2, \dots, f_{Q_0}\}$. Set \mathcal{F} forms level 0 of the hierarchy. These factors are grouped into Q_1 categories of the first level, named $\mathcal{C}^{(1)} = \{c_1^{(1)}, c_2^{(1)}, \dots, c_{Q_1}^{(1)}\}$. Analogously, the first-level categories can be grouped into second-level categories forming a set $\mathcal{C}^{(2)} = \{c_1^{(2)}, c_2^{(2)}, \dots, c_{Q_2}^{(2)}\}$, etc. The entire structure contains L levels. In the last (L -th) level, there is only a single category ($c_1^{(L)}$), called a *root*. Such a structure is presented in Figure 3.2.

When looking at the hierarchy of indicators from the mathematical viewpoint, the factors and categories form a tree. The set of all nodes within this tree is denoted by $\mathcal{N} = \mathcal{F} \cup \mathcal{C}^{(1)} \cup \mathcal{C}^{(2)} \cup \dots \cup \mathcal{C}^{(L)}$. For each node t , except the *root*, we can determine a parent $p(t)$ as a category in which it is directly contained. For each category at hierarchy level $c_i^{(l)}$, we define set $A_{c_i^{(l)}}$ as a subset of \mathcal{F} (inputs and outputs), which are the indirect children of $c_i^{(l)}$. For an elementary factor f , A_f is a singleton, i.e., $A_f = \{f\}$, $f \in \mathcal{F}$. To maintain the spirit of DEA, for each category $c_i^{(l)}$, the respective set of factors ($A_{c_i^{(l)}}$) needs to contain at least one input and one output, i.e., $A_{c_i^{(l)}} \cap \mathcal{IN} \neq \emptyset$ and $A_{c_i^{(l)}} \cap \mathcal{OUT} \neq \emptyset$, for $l = 1, 2, \dots, L$, $i = 1, 2, \dots, Q_l$.

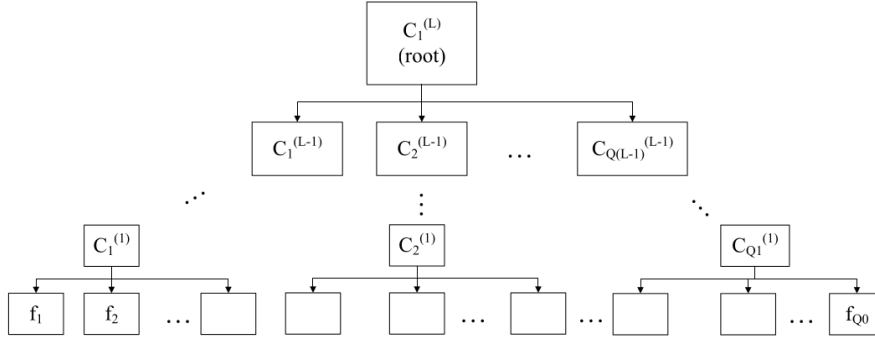


Figure 3.2: A hierarchical structure of inputs and outputs

For each node t , except the root, we assign weight w_t . Moreover, the linear constraints for these weights can be specified at each hierarchy level. Factors or categories involved in the same constraint must have a common parent.

Mathematical programming methods

For the mathematical programming methods, we introduce the additional variables (\hat{w}_q), which allow us to represent the weight constraints without losing the linearity of the problem. For an elementary factor $q \in \mathcal{F}$, the variable \hat{w}_q represents the aggregated weights of q in the hierarchy and is defined as the product of all weights on the path from the analyzed category ($c_i^{(l)}$) at the hierarchy level l to the analyzed factor q :

$$\hat{w}_q^{c_i^{(l)}} = w_q \cdot \prod_{t=1, \dots, l-1 \wedge t \in c_i^{(l)} \wedge q \in A_t} w_t. \quad (3.17)$$

The efficiency of DMU_o is analyzed separately in each node of the hierarchy. For category $c_i^{(l)}$, such an efficiency is defined as follows:

$$E_o^{c_i^{(l)}} = \sum_{q \in A_{c_i^{(l)}}} \hat{w}_q^{c_i^{(l)}} \cdot u_q(DMU_o). \quad (3.18)$$

The true weights assigned to each hierarchy category $c_i^{(l)}$ are defined as the ratio of the sum of weights of indicators contained in this category and the sum of weights of

indicators in the parent category:

$$w_{c_i^{(l)}} = \frac{\sum_{f \in A_{c_i^{(l)}}} \hat{w}_f^{c_i^{(k)}}}{\sum_{f_p \in A_{p(c_i^{(l)})}} \hat{w}_{f_p}^{c_i^{(k)}}}. \quad (3.19)$$

The minimal distance of the analyzed DMU_o to the best unit, when considering the category $c_i^{(l)}$ is determined by solving the following model:

$$\begin{aligned} \min & d_o^{c_i^{(l)}} \\ \text{s.t.} & \sum_{q \in A_{c_i^{(l)}}} \hat{w}_q^{c_i^{(l)}} u_q(DMU_k) - \sum_{q \in A_{c_i^{(l)}}} \hat{w}_q^{c_i^{(l)}} u_q(DMU_o) \leq d_o^{c_i^{(l)}}, \quad \text{for } k = 1, \dots, K, \\ & d_o^{c_i^{(l)}} \geq 0, \\ & \sum_{q \in A_{c_i^{(l)}}} \hat{w}_q^{c_i^{(l)}} = 1, \\ & \hat{w}_q \geq 0, \quad q \in A_{c_i^{(l)}}, \\ & w_t = \frac{\sum_{f \in A_t} \hat{w}_f^{c_i^{(l)}}}{\sum_{f \in A_{p(t)}} \hat{w}_f^{c_i^{(l)}}} \in S_w, \quad \text{for } t \in \mathcal{N} \setminus \{root\}. \end{aligned} \quad (3.20)$$

The remaining LP models, allowing us to determine the maximal distance to the best DMU, the extreme ranks, and the truth of the necessary and possible efficiency preference relations are given in Publication P7.

The robust results for the problems with hierarchical structure have the following properties First, if, for some category, the minimal distance to the best unit is equal to zero in all children nodes, then the minimal distance in this category must also be zero:

Proposition 1 For $DMU_o \in \mathcal{D}$ and category $c_i^{(l)} \in \mathcal{N} \setminus \mathcal{F}$, if $\forall t \in ch(c_i^{(l)}) : d_{o,*}^t = 0$, then $d_{o,*}^{c_i^{(l)}} = 0$.

Similarly, if the minimal rank of some DMU_o is first (or last) in all children categories of some category $c^{(l)}_i$, then its minimal rank must be equal to 1 (K) also in $c^{(l)}_i$:

Proposition 2 For DMU_o and category $c_i^{(l)} \in \mathcal{N} \setminus \mathcal{F}$, if $\forall t \in ch(c_i^{(l)}) : R_{*,o}^t = 1$ then $R_{*,o}^{c_i^{(l)}} = 1$,

Proposition 3 For DMU_o and category $c_i^{(l)} \in \mathcal{N} \setminus \mathcal{F}$, if $\forall t \in ch(c_i^{(l)}) : R_{o,*}^t = K$ then $R_{o,*}^{c_i^{(l)}} = K$.

Moreover, if the maximal rank of DMU_o in all children nodes of some more general category is 1, then it needs to be ranked first in this category even in the worst case:

Proposition 4 For DMU_o and category $c_i^{(l)} \in \mathcal{N} \setminus \mathcal{F}$, if $\forall t \in ch(c_i^{(l)}) : R_o^{*,t} = 1$, then $R_o^{*,c_i^{(l)}} = 1$.

When DMU_o is necessarily preferred to DMU_k in all children nodes of some more general category, then DMU_o must be necessarily preferred to DMU_k given this category:

Proposition 5 For pair (DMU_o, DMU_k) and category $c_i^{(l)} \in \mathcal{N} \setminus \mathcal{F}$, if $\forall t \in ch(c_i^{(l)}) : DMU_o \succsim_E^{N,t} DMU_k$, then $DMU_o \succsim_E^{N,c_i^{(l)}} DMU_k$.

When DMU_o is not possibly preferred to DMU_k in all children nodes of some more general category, then DMU_o is not possibly preferred to DMU_k given this category:

Proposition 6 For pair (DMU_o, DMU_k) and category $c_i^{(l)} \in \mathcal{N} \setminus \mathcal{F}$, if $\forall t \in ch(c_i^{(l)}) : \neg(DMU_o \succsim_E^{P,t} DMU_k)$, then $\neg(DMU_o \succsim_E^{P,c_i^{(l)}} DMU_k)$.

Also, when DMU_o is necessarily preferred to DMU_k in all children nodes of some more general category except one node for which it is possibly preferred, then DMU_o needs to be possibly preferred to DMU_k given this category:

Proposition 7 For pair (DMU_o, DMU_k) and category $c_i^{(l)} \in \mathcal{N} \setminus \mathcal{F}$, if $\forall t \in ch(c_i^{(l)}) \setminus a : DMU_o \succsim_E^{N,t} DMU_k \wedge DMU_o \not\succsim_E^{P,a} DMU_k$, then $DMU_o \succsim_E^{P,c_i^{(l)}} DMU_k$.

Simulation-based methods

We again apply the Monte Carlo simulation to compute the stochastic indices for problems with the hierarchical structure of factors. Similarly to the problem with a flat structure, we estimate the DAIs, ERAIs, and PEOIs as well at the expected distance to the best DMU and the expected rank. However, in the considered situation, we generate weights for all categories and factors, which respect the condition that the sum of weights of categories or factors with the same parent must be equal to one. Moreover, we ensure that the provided weight constraints for all hierarchy levels are satisfied.

3.10 Robustness analysis methods for multiple scenarios of efficiency evaluation

In this section, we describe the extension of the robustness analysis framework proposed in Publication P4. It applies when the same set of DMUs is evaluated under different scenarios. Let us denote a set of such scenarios as \mathcal{S} . For each scenario, the same DMU attains different input and output values, leading to different efficiency results. For example, for the case study considered in the described publication, consisting of evaluating the efficiency of physicians (see Section 4.4), the different scenarios represent the different patients' complaint groups. The proposed extension is an adaptation of the approach proposed for the group decision-making problems [40]. We focus on the exact methods of the robustness analysis framework. Let us denote the results of the analysis of the single scenario ($S \in \mathcal{S}$) as follows: the extreme distances to the best unit as $[d_{*,o,S}, d_{o,S}^*]$, extreme ranks as $[R_{*,o,S}, R_{o,S}^*]$ and the pairwise preference relations as

$\succsim_{E,S}^N$ and $\succsim_{E,S}^P$. Having obtained the robust results for individual scenarios, we consider some robust measures for multiple scenarios as the necessary and possible outcomes representing the results obtained, respectively, for all and at least one scenario. For the necessary efficiency preferences and efficiency ranks, such methods are defined as follows:

- the necessary-necessary preference relation $\succsim_{E,S}^{N,N}$ holds for (DMU_o, DMU_k) if $\forall S \in \mathcal{S} DMU_o \succsim_{E,S}^N DMU_k$;
- the necessary-possible preference relation $\succsim_{E,S}^{N,P}$ holds for (DMU_o, DMU_k) if $\exists S \in \mathcal{S} DMU_o \succsim_{E,S}^N DMU_k$;
- the set of possible-necessary efficiency ranks $[R_{*,o,S}^N, R_{o,S}^{*,N}]$ is a set of ranks attained for all $S \in \mathcal{S}$, i.e., $[R_{*,o,S}^N, R_{o,S}^{*,N}] = \bigcap_{S \in \mathcal{S}} [R_{*,o,S}, R_{o,S}^*]$;
- the set of possible-possible efficiency ranks $[R_{*,o,S}^P, R_{o,S}^{*,P}]$ is a set of ranks attained for at least one $S \in \mathcal{S}$, i.e., $[R_{*,o,S}^P, R_{o,S}^{*,P}] = \bigcup_{S \in \mathcal{S}} [R_{*,o,S}, R_{o,S}^*]$.

The analogous measures can be defined for the other analysis perspectives: the possible preference relations, the efficiency scores, and the distances to the best DMU.

Chapter 4

Extensions and applications

4.1 Selection of a common vector of weights based on the outcomes of robustness analysis

In the traditional DEA approach, we select the different, most favorable, weight vector for each DMU. Such an approach may prevent a justifiable ranking or selection of the best DMU because of the lack of a common base for their comparison [19]. This is why, for some applications, it may be appropriate to determine the common set of weights for all DMUs, giving the joint base for their comparison. The idea of finding the common set of weights in DEA was first introduced in [17]. Over the last years, multiple such methods were proposed [19].

In this section, we describe the novel method of finding the common set of weights based on the outcomes from the robustness analysis proposed in Publication P4. The main idea of the method is to find a single weight vector that represents the most the whole set of feasible input/output weight vectors. Specifically, if the robustness analysis outcomes conclude that DMU_o is better than DMU_k , then the difference between the efficiency scores of them should be enhanced. On the contrary, if the results of the robustness analysis indicate some ambiguity in the comparison of some DMUs, then the difference in their efficiency scores should be as small as possible. The results of the robustness analysis allow us to provide multiple robust relations which confirm the evident advantage of one DMU over the other one. We denote such relation as \succ^W . The incomparability between a pair of DMUs in terms of the relation \succ^w is denoted as R^W . In the publication, we propose four such relations, which are based on the necessary preference relation (\succ_E^N), expected efficiency scores (EEs), expected rank (ERs), and PEOIs. The conditions needed for establishing relations \succ^W and R^W are defined in Table 4.1. For example, when referring to the EEs , one DMU can be judged as preferred to another if the difference between their expected efficiency scores is greater than the predefined threshold t_{EE} . If the absolute value of such difference is lesser than the threshold, then we may assume that the difference is negligible (DMUs are incomparable). A similar approach can be applied for the expected efficiency ranks (ERs) and PEOIs. For the necessary preference relations, we judge one unit better than another if it is necessarily preferred to it. If none of the pair of DMUs is necessarily preferred to another, then we judge them as incomparable.

Having determined the sets of pairs of DMUs for which the relations \succ^W and R^W

Table 4.1: Conditions justifying the truth of the robust preference \succ^W and incomparability R^W relations.

Result	$DMU_o \succ^W DMU_k$	$DMU_l R^W DMU_p$
\succ_E^N	$DMU_o \succ_E^N DMU_k$ and $not(DMU_k \succ_E^N DMU_o)$	$not(DMU_l \succ_E^N DMU_p)$ and $not(DMU_p \succ_E^N DMU_l)$
EE	$EE(DMU_o) - EE(DMU_k) > t_{EE}$	$ EE(DMU_o) - EE(DMU_k) \leq t_{EE}$
ER	$ER(DMU_o) - ER(DMU_k) > t_{ER}$	$ ER(DMU_o) - ER(DMU_k) \leq t_{ER}$
$PEOI$	$PEOI(DMU_o, DMU_k) - PEOI(DMU_k, DMU_o) > t_{PEOI}$	$ PEOI(DMU_l, DMU_p) - PEOI(DMU_p, DMU_l) \leq t_{PEOI}$

hold, selecting the common set of weights is conducted in a two-step procedure. Firstly, we maximize the minimal difference between the efficiency scores for pairs of units related by \succ^W , that is,

$$\begin{aligned}
 & \max \alpha \\
 & \text{subject to:} \\
 & \text{for } (DMU_o, DMU_k), \text{ such that } DMU_o \succ^W DMU_k : \\
 & \sum_{q=1}^Q u_q(DMU_o) - \sum_{q=1}^Q u_q(DMU_k) \geq \alpha.
 \end{aligned}$$

Secondly, we minimize the maximal distance for pairs of units (DMU_l, DMU_p) , for which $DMU_l R^W DMU_p$ (α^* denotes the optimal solution from the previous step):

$$\begin{aligned}
 & \min \beta \\
 & \text{subject to:} \\
 & \text{for } (DMU_l, DMU_p), \text{ such that } DMU_l R^W DMU_p : \\
 & \sum_{q=1}^Q U_q(DMU_l) - \sum_{q=1}^Q u_q(DMU_p) \leq \beta, \\
 & \sum_{q=1}^Q U_q(DMU_p) - \sum_{q=1}^Q u_q(DMU_p s) \leq \beta, \\
 & \text{for } (DMU_o, DMU_k), \text{ such that } DMU_o \succ^W DMU_k : \\
 & \sum_{q=1}^Q u_q(DMU_o) - \sum_{q=1}^Q u_q(DMU_k) \geq \alpha^*, \\
 & \mathcal{W}.
 \end{aligned}$$

4.2 Efficiency reducts and constructs

In Publication P2, we introduce two novel concepts, which aid in generating the explanations of the outcomes of the DEA method:

- the efficiency reduct, for an efficient DMU_o , is a minimal subset of indicators that make it efficient;
- the efficiency construct, for an inefficient DMU_o , is a smallest subset of DMUs, that underlie its inefficiency.

To identify all efficiency reducts, we propose an additive method (see Algorithm 1), which verifies, progressively, the efficiency score of DMU_o using different subsets of inputs and outputs starting with the smallest ones. If for some subset of indicators SUB_k , the DMU_o is efficient, then all supersets of SUB_k are eliminated from further consideration [43]. For each efficient unit, there is at least one efficiency reduct. The analogous algorithm can be applied to the value-based efficiency model.

Algorithm 1 Additive method for identifying all efficiency reducts

Require: sets of inputs \mathcal{IN} and outputs \mathcal{OUT}

Ensure: \mathcal{RED} , all efficiency reducts for DMU_o

$SUB \leftarrow$ all subsets containing at least one input from \mathcal{IN} and one output from \mathcal{OUT}

for each $SUB_k \in SUB$ **do**

 Determine the efficiency score (E_o^*) for DMU_o , by solving the Equation 2.4, with inputs and outputs reduced to SUB_k

if $E_o^* = 1$

$\mathcal{RED} = \mathcal{RED} \cup SUB_k$

 Remove all supersets of SUB_k from SUB

end if

end for

To determine the efficiency constructs for DMU_o , we solve the MILP model constructed for finding the minimal efficiency rank (see Equation 3.7). The DMUs for which the variable b_k is equal to one need to be eliminated from the data set to make DMU_o efficient. The optimal solution of the LP model indicates one of the efficiency constructs: $IC_w = \{DMU_k \in \mathcal{D} : b_k^* = 1\}$. It is possible to determine other constructs by adding the constraints which forbid finding the solutions found in the previous iterations: $(w, w - 1, \dots, 1): \sum_{DMU_k \in IC_w} b_k \leq R_k^* - 2$ [44].

4.3 Experimental comparison of ranking methods in Data Envelopment Analysis

Publication P3 describes the review and experimental comparison of the methods proving the full ranking of DMUs in a context of DEA. We consider fifteen ranking procedures from different categories, such as super- and cross-efficiency, multivariate statistics, decision analysis, benchmarking, virtual DMU, and social networks. The detailed list of the considered procedures is presented in Table 4.2. Four of them are based on the outcomes of the robustness analysis and were originally proposed in this dissertation (Publications P1 and P3).

The four ranking procedures introduced in this dissertation are based on the outcomes from the robustness analysis. The first of them orders the DMUs based on the expected efficiency score (see Section 3.2), i.e., the unit with the greatest expected efficiency score is ranked at the top, etc. Similarly, we construct a ranking based on the expected efficiency rank (see Section 3.3). This time, the DMU with the lowest expected rank is deemed the best, while the one with the greatest expected rank is ranked at the bottom.

The remaining two methods are inspired by MCDA methods and exploit the matrix of PEOIs (see Section 3.4). First, we propose the NFS-PEOI method, which is an adaptation of the Net Flow Score (NFS) procedure used, e.g., in the PROMETHEE methods [13].

In this method, for each DMU_o , we calculate its net flow (Φ_o), which is the difference between the positive (Φ_o^+) and the negative (Φ_o^-) flows. The positive flow quantifies the relative strength of the examined unit, i.e., its advantage over the remaining DMUs. Analogously, the negative flow represents the relative weakness of DMU_o . The ranking of DMUs is based on the overall measure NFS_o defined as follows:

$$NFS_o = \sum_{k=1}^K [PEOI(DMU_o, DMU_k) - PEOI(DMU_k, DMU_o)]. \quad (4.1)$$

The units with the highest NFS_o are the most preferred.

The last ranking procedure, based on the robustness analysis, exploits the PEOIs matrix using the eigenvector method [68]. In this approach, the units are ranked according to their priorities corresponding to the values in the principal eigenvector of the PEOI matrix.

In the discussed publication, we first describe the considered methods and illustrate them with a common small example. Then, we identify and present the features of each ranking procedure, summarizing them as a list of strengths and weaknesses (see Table 4.3), which can be treated as a guide supporting selecting a ranking method for a particular problem. Finally, the rankings provided with different procedures have been compared with five measures [47]:

- Hit Ratio (HR) [12] – a binary measure which is equal to one if both methods rank the same DMU at the top, i.e.

$$HR(R_1, R_2) = \begin{cases} 1, & \text{if } R_1(1) \cap R_2(1) \neq \emptyset, \\ 0, & \text{otherwise,} \end{cases} \quad (4.2)$$

where $R_x(r)$ is a set of DMUs ranked r -th by procedure R_x .

- Normalized Hit Ratio (NHR) [47] – an extension of the HR measure considering the partial agreement between rankings. It is defined as follows:

$$NHR(R_1, R_2) = \frac{R_1(1) \cap R_2(1)}{R_1(1) \cup R_2(1)}. \quad (4.3)$$

- Kendall's τ [49] is determined based on the agreements and disagreements for pairs of DMUs. Firstly, the preferences (\succ^R) and indifferences (\sim^R) observed for pairs of DMUs (DMU_o, DMU_k), within the ranking provided by procedure R, are translated into the numerical values ($p(R, DMU_o, DMU_k)$) as follows:

$$p(R, DMU_o, DMU_k) = \begin{cases} 1, & \text{if } DMU_o \succ^R DMU_k, \\ 0.5, & \text{if } DMU_o \sim^R DMU_k \\ 0, & \text{if } o = k \vee DMU_k \succ^R DMU_o. \end{cases} \quad (4.4)$$

Kendall's τ is then calculated with the following formula:

$$\tau(R_1, R_2, K) = 1 - 4 \frac{d_k(R_1, R_2)}{K \cdot (K - 1)}, \quad (4.5)$$

where $d_k(R_1, R_2)$ is the Kendall's distance:

$$d_k(R_1, R_2) = 0.5 \sum_{DMU_o, DMU_k \text{ in } \mathcal{D} \times \mathcal{D}} |p(R_1, DMU_o, DMU_k) - p(R_2, DMU_o, DMU_k)|. \quad (4.6)$$

- RDM [47] is computed based on the difference between positions attained by the same DMU in rankings provided by different procedures. Let us define the best rank of DMU_o in the ranking provided by R ($r^*(R, DMU_o)$) as the number of other DMUs which are preferred to DMU_o in the ranking R , increased by one. Analogously, $r_*(R, DMU_o)$ denotes the worst possible rank for DMU_o in R and is calculated as the number of other DMUs over which DMU_o is preferred or indifferent, increased by one. The RDM is computed with the following formula:

$$RDM(R_1, R_2, K) = 1 - \frac{\sum_{DMU_o \in \mathcal{D}} |\bar{r}(R_1, DMU_o) - \bar{r}(R_2, DMU_o)|}{max_{\text{diff}}^{\text{rank}}(K)}, \quad (4.7)$$

where $|\bar{r}(R_1, DMU_o) - \bar{r}(R_2, DMU_o)|$ is an average distance between positions attained by DMU_o in the two rankings, R_1 and R_2 :

$$|\bar{r}(R_1, DMU_o) - \bar{r}(R_2, DMU_o)| = \frac{\sum_{r_1=r^*(R_1, DMU_o)}^{r_*(R_1, DMU_o)} \sum_{r_2=r^*(R_2, DMU_o)}^{r_*(R_2, DMU_o)} |r_1 - r_2|}{(r_*(R_1, DMU_o) - r^*(R_1, DMU_o) + 1) \cdot (r_*(R_2, DMU_o) - r^*(R_2, DMU_o) + 1)} \quad (4.8)$$

and

$$max_{\text{diff}}^{\text{rank}}(K) = \begin{cases} \lfloor \frac{K}{2} \rfloor \cdot K, & \text{if } K \text{ is even,} \\ \lceil \frac{K}{2} \rceil \cdot (K - 1), & \text{if } K \text{ is odd.} \end{cases} \quad (4.9)$$

- RAM [47] quantifies how often the same DMU attains the same rank in two rankings. It is a generalization of the NHR measure to the entire ranking:

$$RAM(R_1, R_2, K) = \frac{1}{K} \sum_{r=1}^K RA(R_1, R_2, r), \quad (4.10)$$

where:

$$RA(R_1, R_2, r) = \frac{|R_1(r) \cap R_2(r)|}{|R_1(r) \cup R_2(r)|}. \quad (4.11)$$

The experiments performed to compare the results of the rankings obtained by different procedures were two-fold. On the one hand, we run the procedures on the 960 randomly generated data sets for 96 different problem settings. The input and output values for artificial data sets were generated from the interval $[0 - 1]$ from both uniform and truncated normal distribution. We also distinguished two sizes of problems: typical ones, with 5 to 30 DMUs, and large instances, with 75 to 100 units. On the other sets, we verified the conclusions obtained with the artificial data sets with those obtained for the ten real-world case studies, which represent the most common application areas of DEA, such as finances, education, transportation, healthcare, farming, and the energy industry.

Table 4.2: Considered DEA ranking procedures.

Category	Procedure	Year
Cross-efficiency	Cross-efficiency (CE) [72]	1986
Super-efficiency	Super-efficiency (SE) [6]	1993
Statistical-based methods	Linear discriminant analysis (LDA) [75]	1994
	Canonical correlation analysis (CCA) [35]	1997
	Discriminant analysis of ratios (DR-DEA) [74]	1998
Benchmark-based methods	Slack-adjusted efficiency ranking (BSA) [82]	1996
	Ranking based on changing the reference set (BCRS) [41]	2007
	Interactive benchmark model (BI) [62]	2009
MCDA-based method	AHP-DEA [76]	2000
Virtual DMU	Virtual DMU method (VDMU) [84]	2006
Network-based DEA	Network-based DEA (NDEA) [61]	2010
Methods based on robustness analysis	Expected Efficiency Score (EE) [45]	2017
	Expected Efficiency Rank (ER) [45]	2017
	Net Flow Score based on PEOI (NFS-PEOI) [53]	2021
	Principal Eigenvector based on PEOI (PEV-PEOI) [53]	2021

The outcomes obtained with all five similarity measures were coherent, i.e., all measures identify the same pairs of methods as providing similar rankings. However, combining all of them was beneficial because of the different interpretations of their results. The first two measures (HR and NHR) focus only on the units ranked first. They are useful in choice problems, i.e., when one wants to choose one, the best, DMU. HR is insensitive to the ties at the top of the ranking. It is equal to one when at least one DMU is ranked at the top in both rankings. In contrast, NHR penalizes the ties. That is why, when comparing to the standard CCR model, for 11 out of 15 ranking procedures, the average HR is equal to one. On the contrary, the NHR for methods compared with CCR is never greater than 0.333. The combination of these two measures provides the information that the majority of procedures rank at the top some efficient units (HR) and allow for discriminating between them (NHR).

When it comes to the three measures quantifying the similarity of the whole rankings (Kendall's τ , RDM, and RAM), the observations about the similarities and dissimilarities between pairs of ranking methods are consistent to a great extent and allow to draw trustworthy conclusions.

The results of the analysis allow us to identify three groups of methods that provide similar rankings. The first group contains the cross-efficiency, the VDMU method, and three procedures based on the robustness analysis: EE, ER, and NFS-PEOI. All these approaches summarize the results obtained with different weight vectors. The second group is formed by SE, BSA, BCRS, BI, and NDEA. The procedures within this group focus on the role of different DMUs as a benchmark. Moreover, all methods from this group, except the BI, discriminate only between the efficient units, while the other ones are ranked according to their standard DEA efficiency score. The last group, containing LDA and DR-DEA, is formed by methods that apply statistical methods to find the set of the common weights for all DMUs.

Three ranking procedures (CCA, AHP-DEA, and PEV-PEOI) do not belong to any of the identified groups. They represent some unique concepts, and we did not find significant similarities for other considered approaches.

The important conclusion derived from the performed experimental analysis is that the selection of the ranking method significantly impacts the obtained results. The procedure can be chosen based on its features (strengths and weaknesses) or the underlying idea. However, it may be worth incorporating a few of them, representing different groups, to obtain rankings of DMUs from different viewpoints.

4.4 Case studies

The proposed framework and its extensions were illustrated with multiple case studies. In this section, we describe them briefly.

Efficiency evaluation of Polish airports. In Publication P1 we analyze the efficiency of 11 Polish airports evaluated with 4 inputs, such as an annual capacity of a terminal, a maximal throughput capacity, a dynamic apron capacity, and a catchment area of the airport. The considered outputs include the passenger traffic and the annual number of movements (landings or takeoffs). We presented and discussed the results of the robustness analysis framework with the ratio-based efficiency model for three different situations. Firstly, we considered the standard problem setting, i.e. taking into account the whole set of airports without the weight constraints. Secondly, some additional linear weight constraints were added. Finally, we identified and eliminated the outlier airports (with super-efficiency greater than 2 [9]). The analysis of all three problem settings proves the usefulness of the framework and illustrates the discriminative value of the weight constraints and the applicability of the robustness analysis to indicate and eliminate outlier DMUs.

Quantifying electricity supply resilience of countries. The second application of the robustness analysis framework with the ratio-based efficiency model considers the electricity supply resilience of 140 countries (see Publication P2). In this application, we consider 12 indicators. As the analyzed study is a general benchmarking problem, where DEA is applied for decision-making, the inputs are the indicators with the negative preference order (to be minimized), and the outputs are the ones with the positive preference (to be maximized) [21]. As a result, we identified 4 inputs and 12 outputs representing both the electricity supply situation within the analyzed countries and their political situation. The analysis of countries' electricity resilience was conducted from a few different perspectives. Firstly, we applied the standard ratio-based DEA model to identify the sets of efficient and inefficient DMUs. Then, we run the algorithm described in Section 4.2 to identify the efficiency reducts for efficient countries and constructs for inefficient ones. In the next step, we identified the benchmark efficient countries (HCUs) for all inefficient ones and the necessary improvements to achieve efficiency in both input-oriented and output-oriented perspectives (see Section 2.1). The following part of the analysis included a discussion of the robust results of the framework presented in this dissertation. Finally, we identified and analyzed three scenarios for future development:

1. Singapore: 8% Solar Photovoltaic Electricity Production – comparison of how the increase in photovoltaic energy would affect the robust results of Singapore's electricity resiliency.

2. Singapore: the changes included in Scenario 1 and the increase of the electricity import and GDP per capita (the improvements assumed to be achieved by 2030) – assessment of how these changes would affect the relative efficiency of Singapore,
3. Japan: Required electricity generation portfolio to make Japan efficient: in this scenario, we searched for the set of improvements that should be implemented for inputs and outputs at the same time to make Japan efficient.

Overall, this study showed that combining the CCR model with the robust efficiency analysis provides a holistic methodology that can be applied to multiple problems from different domains.

Efficiency assessment of Emergency Department physicians. Another case study is described in Publication P4 and considers the efficiency assessment of 20 Emergency Department physicians from the Children’s Hospital of Eastern Ontario in Ottawa using the VDEA efficiency model. The considered data set contained three inputs and one output. The physicians’ performances were evaluated separately for different patients’ complaint groups. The primary focus of the case study was on the patients complaining about abdominal pain and constipation ($G1$). The two other complaint groups were considered as separate scenarios of the analysis: fever ($G2$) and lower or upper extremity injury, head injury, and laceration/puncture ($G3$). The paper discusses the full results of the robustness analysis for group $G1$. Moreover, we applied the method described in Section 4.1 to determine the sets of the common weights based on the outcomes from the robustness analysis, specifically the expected efficiencies, expected ranks, the necessary preference relations and PEOIs. Such an approach allowed us to construct 4 rankings of physicians. The comparison of these rankings with Kandall’s τ coefficient [49, 53] showed the high similarity of all these rankings. It leads to the conclusion that different perspectives of the efficiency analysis identify the same physicians as the best, medium, and worst performers. Finally, we aggregated the robust results for different complaint groups using the multi-scenario robustness analysis (see Section 3.10).

Evaluation of Chinese ports and industrial robots. The robustness analysis framework for problems with imprecise information using the ratio-based model was illustrated with two case studies concerning 27 industrial robots and 17 Chinese ports. They are built on the data from, respectively, [69] and [42]. The industrial robots are described in terms of two inputs (cost and vendor reputation) and two outputs (load capacity and velocity). Vendor reputation is treated as an ordinal input, while the load capacity is provided as interval. In the second case study, concerning the Chinese ports, we consider two precise inputs (labor population and energy consumption) and two desirable outputs (cargo throughput – precise and employee satisfaction – ordinal). There is also one undesirable output (water pollutants – interval), which is treated as input, following the [42]. In both case studies, we introduce some weight constraints that prevent individual factors’ overwhelming role. We computed and discussed the robust outcomes for both case studies.

Evaluation of Special Economic Zones in Poland. In Publication P6 we apply the robustness analysis for Imprecise VDEA model (see Section 3.8) to evaluate the performance of Polish Special Economic Zones. We consider 14 zones that are described

with two inputs (the total area and capital expenditures) and two outputs (the number of jobs and financial results). The area and number of jobs are considered interval factors. The interval performances are constructed with the extreme values observed in the analyzed term. Additionally, for each input and output, we account for the admissible range of the marginal value functions. The discussed results demonstrate the usefulness of the outcomes of the robustness analysis.

Hierarchical efficiency analysis of Polish voivodeships' healthcare systems.

The last case study illustrates the usage of the robustness analysis methods for the problem with a hierarchical structure of inputs and outputs (see Section 3.9). In Publication P7 we consider the problem of the efficiency evaluation of healthcare systems in 16 Polish voivodeships. The assessment was conducted from the perspective of the comprehensive efficiency index, including all nine indicators and three sub-problems representing inhabitants' health improvement perspective, efficient financial management, and consumer satisfaction. We discuss the robustness analysis results for each category. The results of all four perspectives allow us to indicate the strong and weak points for individual voivodeships. Moreover, the rankings, based on the robust outcomes (expected distance to the best voivodeship and expected efficiency rank) are compared to those provided by traditional methods, i.e., cross-efficiency and super-efficiency.

Table 4.3: Main advantages and disadvantages of the considered ranking methods.

Method	Advantages	Disadvantages
CE	Multiple weight vectors considered. Drops unrealistic weight schemes. Applies peer and unbiased self-evaluation.	Limited set of common weights. Ambiguity in the selection of weights.
SE	Simplicity. Detecting outliers.	Ranks only efficient units. Occasional infeasibility. No common basis for the comparison of units. Can favor specialized units.
CCA	Common basis for the comparison of units. Ranks all units.	Reliance on a single weight vector. Inefficient unit can be ranked at the top. Occasional infeasibility. Complex application. Sensitivity of eigenvector computation.
LDA	Common basis for the comparison of units. Ranks all units.	Reliance on a single weight vector. Inefficient unit can be ranked at the top. Occasional infeasibility.
DR-DEA	Common basis for the comparison of units. Ranks all units.	Reliance on a single weight vector. Inefficient unit can be ranked at the top.
BSA	Investigates the impact of efficient units on the inefficient ones.	Ranks only efficient units. Complex interpretation of scores.
BCRS	Investigates the impact of efficient units on the inefficient ones. Multiple weight vectors considered. Simple and direct application.	Ranks only efficient units. Limited set of common weights considered.
BI	Investigates the impact of units on the efficiency of others. Ranks all units.	No common basis for the comparison of units.
AHP-DEA	Incorporates DMUs' cross-efficiency comparisons. Ranks all units.	Inefficient unit can be ranked at the top. Sensitivity of eigenvector computation.
NDEA	Considers multiple input-output settings. Ranks all units.	High time complexity. Sensitivity of eigenvector computation. Complex interpretation of scores.
EE	All feasible weight vectors considered. Avoids arbitrary selection of weights. Intuitive interpretation of scores. Ranks all units.	Requires sampling procedure. Inefficient unit can be ranked at the top.
ER	All feasible weight vectors considered. Avoids arbitrary selection of weights. Intuitive interpretation of scores. Ranks all units.	Requires sampling procedure. Averages ordinal measures (ranks). Inefficient unit can be ranked at the top.
PEV-PEOI	All feasible weight vectors considered. Avoids arbitrary selection of weights. Based on DMUs' pairwise comparisons. Ranks all units.	High time complexity. Requires sampling procedure. Sensitivity of eigenvector computation. Inefficient unit can be ranked at the top.
NFS-PEOI	All feasible weight vectors considered. Avoids arbitrary selection of weights. Based on DMUs' pairwise comparisons. Ranks all units.	Requires sampling procedure. Inefficient unit can be ranked at the top.
VDMU	Simplicity and intuitiveness. Low time complexity. Ranks all units.	Changes the original set of DMUs. High sensitivity to outlying DMUs due to incorporating extreme units. Limited set of weight vectors. Ambiguity in the selection of weights.

Chapter 5

Summary

DEA is a tool that evaluates the relative efficiency of DMUs consuming multiple inputs and producing multiple outputs. Its original statement defines the efficiency of a unit as the ratio of the weighted sum of outputs and the weighted sum of inputs. The two formulations of the mathematical programming models represent the equivalent perspectives of productivity, i.e., the perspective of the efficiency scores and the perspective of the unit combinations. The former provides the vector of input and output weights, which gives the best possible efficiency score of the considered DMU. The optimal solution of the latter is the combination of the existing units, which is the projection of a given DMU on the efficient frontier.

Over the decades, multiple researchers worked on developing DEA. They proposed various efficiency models which extend the original CCR model, such as the variable return to scale (BCC) model, the additive model, or the value-based additive (VDEA) efficiency model.

The outcomes of the standard DEA approach provide only information about the most favorable scenario for the examined unit. This dissertation fills the research gap by focusing on the whole spectrum of feasible weight vectors. We propose the robustness analysis framework, which provides two complementary types of results. On the one hand, the mathematical programming models were implemented to determine the exact, extreme outcomes. On the other hand, we use the Monte Carlo simulation to calculate the stochastic indices representing the distributions of the measures over the feasible weight vector space. On the one hand, the extreme results obtained by mathematical programming allow us to investigate the efficiency of units under the most and the least favorable scenario. However, these extreme values are often insufficiently conclusive. The stochastic analysis provides additional information on how efficiencies, ranks, and pairwise preferences are distributed over the feasible weight vector space. On the other hand, we should not focus only on the stochastic indices because of their probabilistic nature. In particular, it is improbable to choose at random the weight vectors corresponding to the extreme values. Thus, it is reasonable to analyze both the exact and stochastic perspectives. We implement the proposed framework for the standard ratio-based efficiency model and the VDEA model.

The proposed robustness analysis methods consider the productivity of DMUs from three different perspectives. First, we consider the efficiency scores and, for the additive model, the distances to the best DMU. The second perspective focuses on the pairwise

comparisons between pairs of units, while the third one on the efficiency ranks. For each of these perspectives, we propose LP models which explore the most and the least favorable scenario for the examined DMU, i.e., the extreme efficiency scores, the extreme distances to the best DMU, the truth of the necessary and the possible preference relations for pairs of units and the extreme efficiency ranks. In addition, we define and compute four stochastic indices calculated using the Hit-And-Run algorithm: Efficiency Acceptability Interval Index, Distance Acceptability Interval Index, Pairwise Efficiency Outranking Index, and Efficiency Rank Acceptability Index. They provide information on how the measures in different perspectives are distributed within the ranges obtained with the exact methods. Finally, we indicate the interdependencies between the robust and stochastic results, the evolution of the outcomes with an incremental specification of weight constraints, and the impact of the outlier removal on the analysis results.

In many real-world problems, it is impossible to collect precise data about DMUs' performances. Such uncertainty results from measurement inaccuracy, cost of such measurement, or changes in input and output values over time. To consider this situation, we adapted the proposed framework to work with the imprecise data in the form of the interval and ordinal inputs and outputs and the admissible ranges of the marginal value functions, defined by two boundary functions. In mathematical programming, we implement the transformation of the interval performances into the precise ones representing the optimistic or pessimistic scenario for the examined unit. Moreover, we introduced the additional variables and constraints, which ensure that the order of DMUs for ordinal factors is remained and that the obtained marginal value functions are monotonic. In addition, for stochastic methods, the sampling procedure needed to be enhanced for problems with imprecise data. We proposed a three-step sampling. First, we obtain the exact performances for units from the given intervals. Next, when VDEA model is applied, we randomly choose the marginal values from the predefined range. Finally, we sample the weight vectors similarly to the standard problems.

This dissertation also considers the efficiency problems with a multiple-layer hierarchical structure of inputs and outputs. In this case, the factors are organized into categories, which can be included in other, more general categories, etc. The main benefits of such a structure are the following. First, it is easy to modify and update the hierarchy. Second, the problem can be decomposed into smaller subproblems, which are manageable and allow to draw more specific conclusions. Third, is it possible to model the interactions at different levels, not only for individual factors. We proposed mathematical programming models, which allow us to find robust outcomes in any category within the hierarchy. Moreover, they consider the weight constraints defined at all hierarchy levels. Similarly, when calculating the acceptability indices, the sampling procedure generates the weights at different hierarchy levels with predefined constraints.

Sometimes, the same set of units is evaluated under different scenarios (e.g., various patients' complaint groups). The input and output values differ for each individual scenario. We propose the methods inspired by the MCDA group decision-making, representing two robustness levels. First, the units are evaluated with the proposed framework separately for each scenario, which allows for analyzing the efficiency within one particular situation. Second, the robust outcomes are aggregated and capture the stability of the results over the different scenarios. We introduce some additional necessary and possible measures which represent the results obtained for, respectively, for all and at least one scenario.

Furthermore, the analysis of the whole set of robust results and acceptability indices may be overwhelming for the DM. That is why we developed a procedure to identify the single, representative weight vector based on the outcomes from the robustness analysis. The approach is based on two goals. On the one hand, for pairs of DMUs, for which the robust outcomes conclude that one unit is preferred to another, we enhance the difference in efficiency scores. On the other hand, for pairs of units that are incomparable according to chosen measure (e.g., the truth of the necessary preference relation), the difference in efficiency scores should be possibly small. Both goals are achieved by solving, in sequence, two linear programming models proposed in this dissertation. Such an approach provides a full ranking of DMU with a common base of comparison while ensuring its maximum possible representation of the whole weight vector space.

The last extension of the robustness analysis framework introduced in this dissertation includes the two algorithms allowing us to find the efficiency reducts and constructs. First of them, i.e., efficiency reducts, are defined for the efficient units and represent the minimal sets of inputs and outputs for which the examined DMU is efficient. We propose an additive algorithm, which identifies all efficiency reducts for some efficient DMU by starting from the smallest sets of factors and progressively verifying the unit's efficiency. The efficiency construct, for an inefficient unit, is a minimal set of other DMUs which make this unit inefficient. We propose the LP model allowing us to find one particular efficiency construct. To determine all of them, the proposed model should be solved multiple times, adding, each time, a constraint that prevents finding the same construct again.

Moreover, we performed the experimental comparison of different ranking procedures for DEA proposed in the literature. We compared fifteen methods, among which four are based on the outcomes of the robustness analysis and were originally proposed in this dissertation. First two methods rank the DMUs based on the expected efficiency scores and the expected ranks obtained with the stochastic robustness analysis. The remaining ones exploit the matrix of PEOIs in two different ways. The method called NFS-PEOI is inspired by the Net Flow Score procedure from MCDA and ranks units according to the difference in their relative strength and weakness gathered from the PEOI matrix. The last method, called PEV-PEOI, ranks the units based on the values in the principal eigenvector of PEOI matrix. The experiments, performed using ten real-world and 960 randomly generated data sets, allowed us to identify three groups of methods providing similar rankings and three methods representing unique concepts for which the provided rankings were not similar to those produced by any other procedure.

The robustness analysis framework and its extension were illustrated with a few case studies, including evaluating airports, Special Economic Zones, and voivodeships' healthcare systems in Poland. Moreover, we extensively analyzed the electricity supply resilience of 140 countries and Emergency Department physicians from the Children's Hospital in Eastern Ontario in Ottawa.

To make the framework presented in this dissertation available, we implemented the proposed methods in R and shared them in the form of modules in an open-source *diviz* platform [63]. Their source code is available at https://github.com/alabijak/diviz_DEA/. Within this dissertation, we created a few tens of modules to compute the robustness analysis outcomes for both ratio-based and VDEA models for standard, i.e., flat and precise problems, problems with imprecise information, and hierarchical ones. The implemented modules can be combined into complex workflows with other modules,

both computational and visualization. Generated workflows can be easily exported and shared with other users. Moreover, *diviz* allows storing the history of workflow executions, which could be useful when comparing the results for different settings.

The dissertation opens some directions for future development. First, the methods for eliciting the shape of the marginal value functions can be adapted and implemented for the proposed framework with the value-based model. As direct elicitation is challenging for the decision-makers, it may be worth indirectly inferring such information, e.g., with pairwise comparisons. Second, the interactions between the factors can be accounted for in the model. Third, the stochastic analysis may consider other probability distributions. Next, the experimental comparison of the ranking methods can be extended to consider other procedures and the possibility of defining some preference information, e.g., weight constraints. Moreover, we could develop the methods for generating explanations of the robust outcomes, similarly to the approaches proposed for the MCDA field (e.g., [37, 51]). The proposed framework should also be adapted for the big data problems addressed in recent DEA studies, e.g., [50, 91]. Finally, we could apply the DEA approach to the sorting problems by classifying the DMUs into multiple preference-ordered efficiency classes.

Bibliography

- [1] N. Adler, L. Friedman, and Z. Sinuany-Stern. Review of ranking methods in the data envelopment analysis context. *European journal of operational research*, 140(2):249–265, 2002.
- [2] T. Ahn, A. Charnes, and W. W. Cooper. Efficiency characterizations in different dea models. *Socio-Economic Planning Sciences*, 22(6):253–257, 1988.
- [3] A. Aldamak and S. Zolfaghari. Review of efficiency ranking methods in data envelopment analysis. *Measurement*, 106:161–172, 2017.
- [4] A. I. Ali, C. S. Lerme, and L. M. Seiford. Components of efficiency evaluation in data envelopment analysis. *European Journal of Operational Research*, 80(3):462–473, 1995.
- [5] A. I. Ali and L. M. Seiford. The mathematical programming approach to efficiency analysis. *The Measurement of Productive Efficiency: Techniques and Applications*, 120:159, 1993.
- [6] P. Andersen and N. C. Petersen. A procedure for ranking efficient units in data envelopment analysis. *Management Science*, 39(10):1261–1264, 1993.
- [7] T. R. Anderson, K. Hollingsworth, and L. Inman. The fixed weighting nature of a cross-evaluation model. *Journal of Productivity Analysis*, 17:249–255, 2002.
- [8] J. Aparicio, J. M. Cordero, and L. Ortiz. Measuring efficiency in education: The influence of imprecision and variability in data on dea estimates. *Socio-Economic Planning Sciences*, 68:100698, 2019.
- [9] R. D. Banker and H. Chang. The super-efficiency procedure for outlier identification, not for ranking efficient units. *European Journal of Operational Research*, 175(2):1311–1320, 2006.
- [10] R. D. Banker, A. Charnes, and W. W. Cooper. Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management Science*, 30(9):1078–1092, 1984.
- [11] R. D. Banker, W. W. Cooper, L. M. Seiford, R. M. Thrall, and J. Zhu. Returns to scale in different dea models. *European Journal of Operational Research*, 154(2):345–362, 2004.
- [12] F. H. Barron and B. E. Barrett. Decision quality using ranked attribute weights. *Management Science*, 42(11):1515–1523, 1996.

- [13] J.-P. Brans and Y. De Smet. *PROMETHEE Methods*, pages 187–219. Springer New York, New York, NY, 2016.
- [14] A. Charnes and W. W. Cooper. Programming with linear fractional functionals. *Naval Research Logistics Quarterly*, 9(3-4):181–186, 1962.
- [15] A. Charnes, W. W. Cooper, B. Golany, L. Seiford, and J. Stutz. Foundations of data envelopment analysis for pareto-koopmans efficient empirical production functions. *Journal of Econometrics*, 30(1-2):91–107, 1985.
- [16] A. Charnes, W. W. Cooper, and E. Rhodes. Measuring the efficiency of decision making units. *European Journal of Operational Research*, 2(6):429–444, 1978.
- [17] A. Charnes, W. W. Cooper, Q. L. Wei, and Z. Huang. Cone ratio data envelopment analysis and multi-objective programming. *International Journal of Systems Science*, 20(7):1099–1118, 1989.
- [18] L. Cherchye, W. Moesen, N. Rogge, T. V. Puyenbroeck, M. Saisana, A. Saltelli, R. Liska, and S. Tarantola. Creating composite indicators with dea and robustness analysis: the case of the technology achievement index. *Journal of the Operational Research Society*, 59(2):239–251, 2008.
- [19] I. Contreras. A review of the literature on dea models under common set of weights. *Journal of Modelling in Management*, 2020.
- [20] W. D. Cook and L. M. Seiford. Data envelopment analysis (dea)—thirty years on. *European Journal of Operational Research*, 192(1):1–17, 2009.
- [21] W. D. Cook, K. Tone, and J. Zhu. Data envelopment analysis: Prior to choosing a model. *Omega*, 44:1–4, 2014.
- [22] W. W. Cooper, K. S. Park, and G. Yu. Idea and ar-idea: Models for dealing with imprecise data in dea. *Management Science*, 45(4):597–607, 1999.
- [23] W. W. Cooper, L. M. Seiford, and K. Tone. *Data envelopment analysis: a comprehensive text with models, applications, references and DEA-solver software*, volume 2. Springer, 2007.
- [24] S. Corrente, S. Greco, and R. Słowiński. Multiple criteria hierarchy process in robust ordinal regression. *Decision Support Systems*, 53(3):660–674, 2012.
- [25] S. Corrente, S. Greco, and R. Słowiński. Handling imprecise evaluations in multiple criteria decision aiding and robust ordinal regression by n-point intervals. *Fuzzy Optimization and Decision Making*, 16:127–157, 2017.
- [26] P. N. de Almeida and L. C. Dias. Value-based dea models: application-driven developments. *Journal of the Operational Research Society*, 63(1):16–27, 2012.
- [27] L. Del Vasto-Terrientes, A. Valls, R. Slowinski, and P. Zielniewicz. Electre-iii-h: An outranking-based decision aiding method for hierarchically structured criteria. *Expert Systems with Applications*, 42(11):4910–4926, 2015.

- [28] D. K. Despotis and Y. G. Smirlis. Data envelopment analysis with imprecise data. *European Journal of Operational Research*, 140(1):24–36, 2002.
- [29] J. Doyle and R. Green. Efficiency and cross-efficiency in dea: Derivations, meanings and uses. *Journal of the Operational Research Society*, 45:567–578, 1994.
- [30] J. Du, L. Liang, and J. Zhu. A slacks-based measure of super-efficiency in data envelopment analysis: A comment. *European Journal of Operational Research*, 204(3):694–697, 2010.
- [31] A. Emrouznejad, B. R. Parker, and G. Tavares. Evaluation of research in efficiency and productivity: A survey and analysis of the first 30 years of scholarly literature in dea. *Socio-Economic Planning Sciences*, 42(3):151–157, 2008.
- [32] A. Emrouznejad and G.-l. Yang. A survey and analysis of the first 40 years of scholarly literature in dea: 1978–2016. *Socio-economic planning sciences*, 61:4–8, 2018.
- [33] M. J. Farrell. The measurement of productive efficiency. *Journal of the Royal Statistical Society*, 120(3):253–281, 1957.
- [34] P. C. Fishburn. Methods of estimating additive utilities. *Management Science*, 13(7):435–453, 1967.
- [35] L. Friedman and Z. Sinuany-Stern. Scaling units via the canonical correlation analysis in the DEA context. *European Journal of Operational Research*, 100(3):629–637, 1997.
- [36] P. Gasser, M. Cinelli, A. Labijak, M. Spada, P. Burgherr, M. Kadziński, and B. Stojadinović. Quantifying electricity supply resilience of countries with robust efficiency analysis. *Energies*, 13(7):1535, 2020.
- [37] J. Geldermann. *Explanation Systems*, pages 241–259. Springer Netherlands, Dordrecht, 2010.
- [38] M. C. Gouveia, L. C. Dias, and C. H. Antunes. Additive dea based on mcda with imprecise information. *Journal of the Operational Research Society*, 59(1):54–63, 2008.
- [39] M. C. Gouveia, L. C. Dias, and C. H. Antunes. Super-efficiency and stability intervals in additive dea. *Journal of the Operational Research Society*, 64(1):86–96, 2013.
- [40] S. Greco, M. Kadziński, V. Mousseau, and R. Słowiński. Robust ordinal regression for multiple criteria group decision: Utagms-group and utadisgms-group. *Decision Support Systems*, 52(3):549–561, 2012.
- [41] G. R. Jahanshahloo, H. V. Junior, F. H. Lotfi, and D. Akbarian. A new DEA ranking system based on changing the reference set. *European Journal of Operational Research*, 181(1):331–337, 2007.
- [42] B. Jiang, C. Yang, Q. Dong, and J. Li. Ecological efficiency evaluation of china’s port industries with imprecise data. *Evolutionary Intelligence*, pages 1–12, 2021.

- [43] M. Kadziński, S. Corrente, S. Greco, and R. Słowiński. Preferential reducts and constructs in robust multiple criteria ranking and sorting. *OR Spectrum*, 36:1021–1053, 2014.
- [44] M. Kadziński, S. Greco, and R. Słowiński. Robust ordinal regression for dominance-based rough set approach to multiple criteria sorting. *Information Sciences*, 283:211–228, 2014.
- [45] M. Kadziński, A. Labijak, and M. Napieraj. Integrated framework for robustness analysis using ratio-based efficiency model with application to evaluation of polish airports. *Omega*, 67:1–18, 2017.
- [46] M. Kadziński and T. Tervonen. Stochastic ordinal regression for multiple criteria sorting problems. *Decision Support Systems*, 55(1):55–66, 2013.
- [47] M. Kadziński and M. Michalski. Scoring procedures for multiple criteria decision aiding with robust and stochastic ordinal regression. *Computers & Operations Research*, 71:54–70, 2016.
- [48] R. L. Keeney and H. Raiffa. *Decisions with multiple objectives: preferences and value trade-offs*. Cambridge university press, 1993.
- [49] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.
- [50] D. Khezrimotlagh, J. Zhu, W. D. Cook, and M. Toloo. Data envelopment analysis and big data. *European Journal of Operational Research*, 274(3):1047–1054, 2019.
- [51] D. A. Klein. *Decision-analytic intelligent systems: automated explanation and knowledge acquisition*. Psychology Press, 1994.
- [52] C. T. Kuah, K. Y. Wong, and F. Behrouzi. A review on data envelopment analysis (dea). In *2010 Fourth Asia International Conference on Mathematical/Analytical Modelling and Computer Simulation*, pages 168–173. IEEE, 2010.
- [53] A. Labijak-Kowalska and M. Kadziński. Experimental comparison of results provided by ranking methods in data envelopment analysis. *Expert Systems with Applications*, 173:114739, 2021.
- [54] A. Labijak-Kowalska and M. Kadziński. Exact and stochastic methods for robustness analysis in the context of imprecise data envelopment analysis. *Operational Research*, 23(1):22, Mar 2023.
- [55] A. Labijak-Kowalska, M. Kadziński, and L. C. Dias. Robustness analysis for imprecise additive value efficiency analysis with an application to evaluation of special economic zones in poland. 2023. Submitted to Socio-Economic Planning Sciences.
- [56] A. Labijak-Kowalska, M. Kadziński, I. Sychała, L. C. Dias, J. Fiallos, J. Patrick, W. Michalowski, and K. Farion. Performance evaluation of emergency department physicians using robust value-based additive efficiency model. *International Transactions in Operational Research*, 30(1):503–544, 2023.

- [57] A. Labijak-Kowalska, M. Kadziński, and W. Mrozek. Robust additive value-based efficiency analysis with a hierarchical structure of inputs and outputs. *Applied Sciences*, 13(11), 2023.
- [58] R. Lahdelma and P. Salminen. Smaa-2: Stochastic multicriteria acceptability analysis for group decision making. *Operations Research*, 49(3):444–454, 2001.
- [59] R. Lahdelma and P. Salminen. Stochastic multicriteria acceptability analysis using the data envelopment model. *European journal of operational research*, 170(1):241–252, 2006.
- [60] J. S. Liu, L. Y. Lu, W.-M. Lu, and B. J. Lin. A survey of dea applications. *Omega*, 41(5):893–902, 2013.
- [61] J. S. Liu and W.-M. Lu. DEA and ranking with the network-based approach: a case of R&D performance. *Omega*, 38(6):453–464, 2010.
- [62] W.-M. Lu and S.-F. Lo. An interactive benchmark model ranking performers - application to financial holding companies. *Mathematical and Computer Modelling*, 49(1-2):172–179, 2009.
- [63] P. Meyer and S. Bigaret. Diviz: A software for modeling, processing and sharing algorithmic workflows in mcd. *Intelligent decision technologies*, 6(4):283–296, 2012.
- [64] R. Pelissari, M. C. Oliveira, S. B. Amor, A. Kandakoglu, and A. L. Helleno. Smaa methods and their applications: a literature review and future research directions. *Annals of Operations Research*, 293:433–493, 2020.
- [65] V. V. Podinovski. The explicit role of weight bounds in models of data envelopment analysis. *Journal of the Operational Research Society*, 56(12):1408–1418, 2005.
- [66] S. C. Ray. *Data envelopment analysis: theory and techniques for economics and operations research*. Cambridge university press, 2004.
- [67] B. Roy. Robustness in operational research and decision aiding: A multi-faceted issue. *European Journal of Operational Research*, 200(3):629–638, 2010.
- [68] T. L. Saaty. What is the analytic hierarchy process? In G. Mitra, H. J. Greenberg, F. A. Lootsma, M. J. Rijkaert, and H. J. Zimmermann, editors, *Mathematical Models for Decision Support*, pages 109–121, Berlin, Heidelberg, 1988. Springer Berlin Heidelberg.
- [69] R. F. Saen. Technologies ranking in the presence of both cardinal and ordinal data. *Applied Mathematics and Computation*, 176(2):476–487, 2006.
- [70] A. Salo and A. Punkka. Ranking intervals and dominance relations for ratio-based efficiency analysis. *Management Science*, 57(1):200–214, 2011.
- [71] L. M. Seiford and J. Zhu. Stability regions for maintaining efficiency in data envelopment analysis. *European Journal of Operational Research*, 108(1):127–139, 1998.
- [72] T. R. Sexton, R. H. Silkman, and A. J. Hogan. Data envelopment analysis: Critique and extensions. *New Directions for Program Evaluation*, 1986(32):73–105, 1986.

- [73] Y. Shen, E. Hermans, D. Ruan, G. Wets, T. Brijs, and K. Vanhoof. A generalized multiple layer data envelopment analysis model for hierarchical structure assessment: A case study in road safety performance evaluation. *Expert systems with applications*, 38(12):15262–15272, 2011.
- [74] Z. Sinuany-Stern and L. Friedman. DEA and the discriminant analysis of ratios for ranking units. *European Journal of Operational Research*, 111(3):470–478, 1998.
- [75] Z. Sinuany-Stern, A. Mehrez, and A. Barboy. Academic departments efficiency via DEA. *Computers & Operations Research*, 21(5):543–556, 1994.
- [76] Z. Sinuany-Stern, A. Mehrez, and Y. Hadad. An AHP/DEA methodology for ranking decision making units. *International Transactions in Operational Research*, 7(2):109–124, 2000.
- [77] T. Sueyoshi, Y. Yuan, and M. Goto. A literature study for dea applied to energy and environment. *Energy Economics*, 62:104–124, 2017.
- [78] T. Tervonen, G. van Valkenhoef, N. Baştürk, and D. Postmus. Hit-and-run enables efficient weight generation for simulation-based multiple criteria decision analysis. *European Journal of Operational Research*, 224(3):552–559, 2013.
- [79] R. G. Thompson, L. N. Langemeier, C.-T. Lee, E. Lee, and R. M. Thrall. The role of multiplier bounds in efficiency analysis with application to kansas farming. *Journal of econometrics*, 46(1-2):93–108, 1990.
- [80] K. Tone. A slacks-based measure of efficiency in data envelopment analysis. *European Journal of Operational Research*, 130(3):498–509, 2001.
- [81] K. Tone. A slacks-based measure of super-efficiency in data envelopment analysis. *European Journal of Operational Research*, 143(1):32–41, 2002.
- [82] A. M. Torgersen, F. R. Førsund, and S. A. Kittelsen. Slack-adjusted efficiency measures and ranking of efficient units. *Journal of Productivity Analysis*, 7(4):379–398, 1996.
- [83] P. Vincke. Robust solutions and methods in decision-aid. *Journal of Multi-criteria Decision Analysis*, 8(3):181–187, 1999.
- [84] Y.-M. Wang and Y. Luo. DEA efficiency assessment using ideal and anti-ideal decision making units. *Applied Mathematics and Computation*, 173(2):902–915, 2006.
- [85] P. W. Wilson. Detecting influential observations in data envelopment analysis. *Journal of Productivity Analysis*, 6(1):27–45, 1995.
- [86] J. Wu, J. Chu, J. Sun, and Q. Zhu. Dea cross-efficiency evaluation based on pareto improvement. *European Journal of Operational Research*, 248(2):571–579, 2016.
- [87] J. Wu, J. Sun, and L. Liang. Dea cross-efficiency aggregation method based upon shannon entropy. *International Journal of Production Research*, 50(23):6726–6736, 2012.

- [88] J. Wu, J. Sun, and L. Liang. Methods and applications of dea cross-efficiency: Review and future perspectives. *Frontiers of Engineering Management*, 8(2):199–211, 2021.
- [89] J. Zhu. Robustness of the efficient dmus in data envelopment analysis. *European Journal of Operational Research*, 90(3):451–460, 1996.
- [90] J. Zhu. Imprecise data envelopment analysis (idea): A review and improvement with an application. *European Journal of Operational Research*, 144(3):513–529, 2003.
- [91] J. Zhu. Dea under big data: Data enabled analytics and network data envelopment analysis. *Annals of Operations Research*, 309(2):761–783, 2022.

Publication reprints

Publication [P1]

M. Kadziński, A. Labijak, and M. Napieraj. Integrated framework for robustness analysis using ratio-based efficiency model with application to evaluation of polish airports. *Omega*, 67:1–18, 2017, DOI: 10.1016/j.omega.2016.03.003.

Number of citations¹:

- according to Web of Science: 24
- according to Google Scholar: 39

Contribution of co-authors (excluding the author and the supervisor of this dissertation):

- Małgorzata Napieraj
 - implementation of the part of the modules of the robustness analysis methods for DEA dedicated to the *diviz* platform,
 - preparation, in tabular form, of some results of the robustness analysis for the problem of the efficiency evaluation of Polish airports.

¹as on May 30, 2023



Integrated framework for robustness analysis using ratio-based efficiency model with application to evaluation of Polish airports[☆]



Miłosz Kadziński^{*}, Anna Labijak, Małgorzata Napieraj

Institute of Computing Science, Poznań University of Technology, Piotrowo 2, 60-965 Poznań, Poland

ARTICLE INFO

Article history:

Received 29 April 2015

Accepted 5 March 2016

Available online 16 March 2016

Keywords:

Data envelopment analysis

Ratio-based efficiency

Robustness analysis

Stochastic multicriteria acceptability analysis

Airport efficiency

Software

ABSTRACT

We consider a problem of evaluating efficiency of Decision Making Units (DMUs) based on their deterministic performance on multiple consumed inputs and multiple produced outputs. We apply a ratio-based efficiency measure, and account for the Decision Maker's preference information representable with linear constraints involving input/output weights. We analyze the set of all feasible weights to answer various robustness concerns by deriving: (1) extreme efficiency scores and (2) extreme efficiency ranks for each DMU, (3) possible and necessary efficiency preference relations for pairs of DMUs, (4) efficiency distribution, (5) efficiency rank acceptability indices, and (6) pairwise efficiency outranking indices. The proposed hybrid approach combines and extends previous results from Ratio-based Efficiency Analysis and the SMAA-D method. The practical managerial implications are derived from the complementary character of accounted perspectives on DMUs' efficiencies. We present an innovative open-source software implementing an integrated framework for robustness analysis using a ratio-based efficiency model on the *diviz* platform. The proposed approach is applied to a real-world problem of evaluating efficiency of Polish airports. We consider four inputs related to the capacities of a terminal, runways, and an apron, and to the airport's catchment area, and two outputs concerning passenger traffic and number of aircraft movements. We present how the results can be affected by integrating the weight constraints and eliminating outlier DMUs.

© 2016 Elsevier Ltd. All rights reserved.

1. Introduction

The framework of Data Envelopment Analysis (DEA) offers a variety of methods for evaluating the relative efficiency of Decision Making Units (DMUs) which consume multiple inputs and produce multiple outputs [18,39,38]. Conceptually, efficiency is the ratio between virtual output and virtual input, i.e., respectively, outputs or inputs aggregated using some weights assigned to these factors [14]. Typically, DEA methods have been used to classify the DMUs into efficient and inefficient ones. By definition, the former ones have an efficiency score equal to one, whereas for the latter ones this measure is less than one. For the inefficient DMUs, such scores convey information on how close to being efficient they are. Analysis of these measures may lead to formulating the corrective actions, revealing an excess use of some inputs or shortfalls in the production of outputs, as well as to indicating a reference set of some comparable DMUs.

1.1. Critical view on the traditional methods of data envelopment analysis

Although DEA has proven its usefulness when applied to a variety of real-world problems (see, e.g., [23,18,40]), some criticism has been leveled against its discriminative power and the way the efficiency scores are computed. Firstly, the efficiency measures for each DMU are derived from the analysis of the input/output weights which are the most favorable to it. However, a weight vector for which a DMU attains its maximal efficiency is not unique [36]. Thus, choosing among them is arbitrary to a large extent. Secondly, the underlying Linear Programming (LP) techniques require some normalization of weights for each DMU individually. This implies that scaling affects the optimal weights and a meaningful comparison of these weights across different DMUs is difficult. Thirdly, the efficiency measures fail to reflect how the efficiencies of DMUs compare to each other for other feasible weight vectors [53]. In fact, only extremely small share of feasible weights is taken into account in the analysis, while others are neglected despite being equally desirable. Fourth, DEA measures efficiency relative to the efficient frontier. This requires some assumptions about possible returns to scale (e.g., constant or variable). These may be, however, difficult to formulate or justify.

[☆]This manuscript was processed by Associate Editor Lim.

^{*} Corresponding author. Tel.: +48 61 665 3022.

E-mail addresses: miłosz.kadziński@cs.put.poznan.pl (M. Kadziński), anna.labijak@student.put.poznan.pl (A. Labijak), napieraj.malgorzata@gmail.com (M. Napieraj).

Further, we may sometimes prefer a DMU judged as inefficient, which is dominated only by some convex combination of other DMUs, but not by any existing DMU [36]. Moreover, an efficiency frontier and, thus, the efficiency scores, vastly depend on the DMUs under consideration [74,58]. The outcomes of DEA may be very sensitive even to the inclusion or removal of a single DMU. In the same spirit, the outcomes of DEA can be interpreted only when the number of DMUs is large enough in comparison with the number of inputs and outputs. Finally, while DEA is useful for indicating which DMUs are efficient, it does not discriminate between them. In some real-world situations, the share of efficient DMUs may be very large, and we may wish to identify among them a small subset of the most distinguishing ones.

Several techniques have been proposed in the literature to address these drawbacks. In particular, preference information on the relative comparisons of inputs and/or outputs may be used to reduce the space of feasible weight vectors [63,50], and, thus, the conclusiveness of efficiency scores. Further, the cross-efficiency methods exploit the space of feasible weights to derive for each DMU an average efficiency obtained from the analysis of weights for which other DMU's efficiency is maximal [59,22]. Moreover, the super-efficiency discriminates the efficient DMUs by indicating for each of them how much more efficient it can be relative to the remaining ones [2,75]. Although following the right direction, these approaches do not address all aforementioned concerns comprehensively. Doing so, requires incorporation of robustness analysis into the DEA framework.

1.2. Existing approaches for robustness analysis in data envelopment analysis

Robustness analysis accounts for the uncertainties which can be observed in the real-world decision problems [33]. A conclusion is considered to be robust if it is true for all or for the most plausible combinations of parameter values [52,67]. As noted in [20], this type of analysis provides information that may allow the users to avoid answering questions they find too demanding. It may also guide them in revising or enriching the provided preference information, progressively constraining the space of admissible values for the parameters of employed model. In the context of DEA, robustness concern refers to the relative efficiencies of DMUs for all feasible input and output weights or their representative sample. Advances in this regard, that we build on in this paper, have been presented in [53] and [36].

On one hand, [53] consider the whole set of weights that are compatible with the preference information concerning input/output variables. The so-called Ratio-based Efficiency Analysis (REA) does not make any assumptions in terms of the production possibilities beyond the set of DMUs that are under comparison. To materialize the relations between the DMUs' efficiencies, the method exhibits three kinds of results derived from the analysis of the whole set of feasible weights: efficiency bounds exposing the greatest and the least relative efficiencies of a DMU compared to a subset of other DMUs, dominance relation indicating for a pair of DMUs if one of them dominates the other in pairwise efficiency comparison, and ranking intervals indicating the range of efficiency ranks that are attained by a DMU. All these results are derived from comparing DMUs' efficiencies pairwise rather than measuring their distance from an efficient frontier as in the traditional DEA models. As a result, these outcomes are interpretable even if the set of DMU is relatively small, being at the same time less sensitive to the inclusion of DMU whose input/output values are distant from the performances of other units.

On the other hand, [36] apply simulation to provide stochastic indices which characterize the possible outcomes of a decision problem. In Stochastic Multicriteria Acceptability Analysis for Data

Envelopment Analysis (SMAA-D), it is possible to handle imprecision and uncertainty regarding the input/output weights and performances of DMUs. The method computes rank acceptability indices which measure the variety of model variables that grant each DMU any rank from the best to the worst. In particular, the best (most acceptable) DMUs are those with high acceptabilities for the first rank. When compared with the basic DEA models, the stochastic measures originally provided in SMAA-D have been found useful for making the efficient DMUs more comparable [36].

1.3. Aim of the paper

The aim of this paper is fourth-fold. Firstly, from a methodological point of view, we extend the range of outcomes considered in REA and SMAA-D. With respect to the robustness analysis, we show how to determine the least efficiency measure for each DMU, i.e., what is the lower bound of the efficiency range when the whole set of DMUs (including the DMU under consideration) is analyzed. When considering stability of the efficiency comparison for pairs of DMUs, we propose to consider the necessary and possible efficiency preference relations instead of the dominance relation. The necessary relation needs to be confirmed by all feasible weight vectors, while the possible one has to be supported by at least one feasible weight vector. We show that taking into account these results is more beneficial than analyzing the dominance relation because of their interpretability and intuitive convergence with the growth of the preference information for input/output variables.

When it comes to SMAA-D, we significantly enrich the range of stochastic indices that can be derived from the representative sample of weight vectors so that they additionally capture the efficiency scores and pairwise efficiency relations. In particular, we analyze the extreme observed efficiencies, the distribution of efficiency measures, and pairwise efficiency outranking (winning) indices indicating the probability that one DMU has an efficiency at least as good (better) than the other. In this way, we provide both exact and stochastic outcomes reflecting three different perspectives on DMUs' efficiency: scores, pairwise preference relations, and attained ranks.

Secondly, we clearly demonstrate the benefits of considering together the outcomes of thus revised REA and SMAA-D. On one hand, with the necessary, possible, and extreme outcomes of the revisited REA, we can analyze what happens for all, some, the most and the least advantageous model parameters. However, the difference between extreme ranks and efficiencies may often be very large, and in practical decision analysis the information on the sole possibility of attaining a particular rank or an efficiency in a given subinterval may be insufficient. Similarly, REA leaves incomparable the pairs of units which are possibly preferred to each other. In this perspective, SMAA-D may enrich REA with answering questions on how probable are the possible efficiency preference relations and what is the distribution of ranks or efficiencies between the best and the worst ones. These results can be further exploited to indicate the expected rank (efficiency) for a given DMU, the ranks (efficiencies) which are attained most often, and the probability of being judged as efficient (obtaining the highest efficiency).

On the other hand, even though the stochastic indices can be estimated with high accuracy using Monte Carlo simulation, they are not exact. In particular, it may be unlikely to hit the weight vector corresponding to the extreme results. This, in turn, implies that such results would not be reflected in the distribution of ranks or efficiency scores. For the same reason, an estimated pairwise efficiency outranking index equal to one or zero does not, respectively, confirm the necessity or exclude the possibility of one DMU being preferred over another. Still, all these input/output

weights whose indications are not reflected in the estimations of stochastic indices are feasible. Thus, it is desirable to confront the indices derived from Monte Carlo simulation with the possible, necessary, and extreme outcomes of exact robustness analysis conducted with LP techniques.

By combining REA and SMAA-D within an integrated framework incorporating robustness and stochastic analysis, we provide a DEA-type variant of hybrid methods that have been recently proposed in Multiple Criteria Decision Aiding (MCDA) [32,33]. In this way, we tighten the interrelations between DEA and MCDA (for a comparison of these two methodological frameworks, see, e.g., [3,13,17,28,29,55,61]).

The third contribution of this paper consists in presenting an open-source software implementing the methods for robustness analysis using ratio-based data envelopment model. They are made available in the form of independent software components on the *diviz* platform [42]. These modules can be subsequently combined, using an intuitive user interface, to construct complex algorithmic workflows. From a technological point of view, they are implemented as web-services, which read the input formatted with respect to a well-defined XML-based standard. The basic components we provide deliver either exact results using GLPK solver or stochastic indices using a Hit-And-Run sampling procedure [62]. Apart from analyzing in this way three types of results concerning efficiency scores, pairwise efficiency preference relations, and efficiency ranks, we enrich the range of DEA-based tools that can be used within *diviz* by providing modules which derive, e.g., cross-efficiency or super-efficiency scores. All implemented components allow incorporating linear weight constraints.

Finally, we apply the presented methodological framework to the real-world problem of evaluating efficiency of Polish airports. We take into account four inputs and two outputs. The inputs are related to the capacities of a terminal, runways, and an apron, and to the airport's catchment area. The outputs concern passenger traffic and number of aircraft movements. By illustrating the use of DEA-based robustness analysis for this particular problem, we prove the usefulness of the proposed approach for studying the performances and measuring the efficiency of airports. This type of research has aroused great interest in the recent years (see, e.g., [26,24,48,25,7,70,72]). Nevertheless, the introduced framework should be perceived as more general one; its use is not limited to this particular domain.

The remainder of this paper is organized as follows. Section 2 presents the new hybrid approach for DEA, combining and extending the ideas from REA and SMAA-D. We present how to compute and interpret robust outcomes and stochastic indices. We also discuss the interdependencies between these two types of results as well as the evolution of robust results with incremental specification of weight constraints. Section 3 concerns an open-source software implementing the proposed integrated framework for robustness analysis in DEA. Section 4 is devoted to the real-world case study investigating efficiency of Polish airports. In Section 5, we focus on the practical considerations. Section 6 concludes the paper.

2. Integrated framework for robustness analysis using ratio-based efficiency measure

2.1. Notation and basic concepts

The following notation is used in the paper:

- $\mathcal{D} = \{DMU_1, \dots, DMU_K\}$ – the set of considered DMUs; thus, K is the number of compared DMUs ($K = |\mathcal{D}|$);
- x_m – m -th input, $m \in \{1, \dots, M\}$;

- y_n – n -th output, $n \in \{1, \dots, N\}$;
- x_{mo} – an amount of m -th input consumed by $DMU_o \in \mathcal{D}$;
- y_{no} – an amount of n -th output produced by $DMU_o \in \mathcal{D}$;
- $v = \{v_1, \dots, v_m\}$ – a vector of input weights;
- $u = \{u_1, \dots, u_n\}$ – a vector of output weights;
- $S_v = \{v = (v_1, \dots, v_M)^T \neq 0 \mid v \geq 0, A_v v \leq 0\}$ and $S_u = \{u = (u_1, \dots, u_N)^T \neq 0 \mid u \geq 0, A_u u \leq 0\}$ – a space of feasible input and output weights, respectively; A_v and A_u are matrices of coefficients derived from linear constraint on weights representing the user's (Decision Maker's) preference information.

To measure the efficiency of each $DMU_o \in \mathcal{D}$, we apply the ratio of virtual output for $u \in S_u$ and virtual input for $v \in S_v$, defined as follows:

$$E_o(v, u) = \frac{\sum_{n=1}^N u_n y_{no}}{\sum_{m=1}^M v_m x_{mo}}. \quad (1)$$

For all feasible weights, the virtual inputs and outputs need to be strictly positive. For conditions satisfying this assumption, see [53].

Referring to the set of feasible weight vectors $(v, u) \in (S_v, S_u)$, robustness of the efficiency analysis may concern three points of view: efficiency scores, pairwise efficiency preference relations, and efficiency ranks. In this section, we discuss in detail two complementary ways for conducting such analysis. On one hand, LP techniques are employed to determine in an exact way: extreme efficiencies and ranks for each DMU as well as verifying the truth of the necessary and possible efficiency preference relations. On the other hand, Monte Carlo simulation algorithms are used to compute stochastic indices based on a representative sample of feasible weight vectors. The latter approach is based on normalizing input and output weights so that the following constraint is respected:

$$\sum_{n=1}^N u_n = \sum_{m=1}^M v_m = 1.$$

This normalization makes the space of feasible weights bounded. Then, a Hit-And-Run method is used to efficiently sample weights from the convex space of feasible weights [62,66]. For this purpose, some probability distribution with joint density function in the feasible weight space needs to be assumed. Such distribution constitutes a form of partial preference information provided by an analyst. In general, our approach can work with any arbitrarily provided distribution. However, in most decision situations, its specification would be rather challenging. Thus, following SMAA-D [36] and MCDA-based Stochastic Ordinal Regression (SOR) [33,32], when other weight distribution is not exogenously given, we use a uniform one. In this way, each weight vector has equal chances ($= 1/\text{vol}(W)$, where $\text{vol}(W)$ is the volume of the feasible weight space) to be considered within a sample of weights. This assumption is also in line with the spirit of robustness analysis, where each individual feasible weight vector is equally authorized to make some outcome non-necessary or possible, or shift the extreme bounds.

For each sampled input/output weight vector, we compute efficiency scores for all DMUs, and then normalize them by the maximal obtained efficiency. In this way, the final efficiency measures are in the interval between zero and one as in the traditional DEA methods. Such results are analyzed to derive estimates of the shares of feasible weight vectors for which: a DMU attains an efficiency score in some pre-defined efficiency sub-interval or a specific rank, and for which some DMU is preferred to another.

When it comes to weight restrictions, as noted in [49], typical examples of such constraints are absolute weight bounds (e.g., $2 \leq v_1 \leq 5$), bounds on virtual inputs or outputs (e.g., $5v_1 + v_2 \geq 1$

or $2u_1 + 3u_2 \leq 1$), and bounds on the ratio of two weights (e.g., $0.5 \leq u_1/u_2 \leq 2 \Rightarrow 0.5u_2 \leq u_1 \leq 2u_2$). All these forms are admitted within the proposed framework.

2.2. Efficiency scores

In this section, we discuss the measures that are useful for analysis of efficiency scores attained by the DMUs across all feasible weight vectors. When compared to REA, we additionally discuss how to determine the lower bound of the efficiency range when the whole set of DMUs (including the DMU under consideration) is analyzed within a single mathematical programming model. When compared to SMAA-D, we propose to consider the efficiency acceptability interval indices which capture the distribution of efficiency scores attained by each DMU.

2.2.1. Extreme efficiency scores

For each $DMU_o \in \mathcal{D}$, the best E_o^* and the worst $E_{o,*}$ efficiencies attained in the set of feasible weight vectors (S_v, S_u) may be computed using LP. The following program needs to be solved to determine E_o^* :

$$\begin{aligned} \max \quad & E_o^* = \sum_{n=1}^N u_n y_{no} \\ \text{subject to:} \quad & \sum_{m=1}^M v_m x_{mo} = 1, \\ & \sum_{n=1}^N u_n y_{nk} \leq \sum_{m=1}^M v_m x_{mk}, \quad k = 1, \dots, K, \quad (v, u) \in (S_v, S_u). \end{aligned} \quad (2)$$

The idea underlying problem (2) consists in finding the most advantageous feasible weight vector $(v, u) \in (S_v, S_u)$ for DMU_o in terms of its efficiency score. Note that E_o^* is equivalent to the efficiency originally proposed in the CCR model [14]. Thus, if $E_o^* = 1$, DMU_o is efficient; otherwise, it is inefficient. The worst efficiency $E_{o,*}$ can be determined with the following LP:

$$\begin{aligned} \min \quad & E_{o,*} = \sum_{n=1}^N u_n y_{no} \\ \text{subject to:} \quad & \sum_{m=1}^M v_m x_{mo} = 1, \\ & \sum_{n=1}^N u_n y_{nk} \geq \sum_{m=1}^M v_m x_{mk} - C(1 - b_k), \quad k = 1, \dots, K, \\ & \sum_{k=1}^K b_k \geq 1, \quad b_k \in \{0, 1\}, \quad k = 1, \dots, K, \quad (v, u) \in (S_v, S_u). \end{aligned} \quad (3)$$

In the above problem, we adapt a more general technique for dealing with inconsistency in LP which is called “The Big-M (or Big-C) method” or “Exact Big-M MIP Formulation” [43,15]. To prevent undesired compensations, this technique assumes that the value assigned to constant C is great enough. In our context, it is sufficient if:

$$C > \max_{DMU_o, DMU_k \in \mathcal{D}} \left\{ \max_{m=1, \dots, M} \{x_{mk}/x_{mo}\} - \min_{n=1, \dots, N} \{y_{nk}/y_{no}\} \right\}.$$

For all values of C satisfying this condition, we are guaranteed to obtain the same results.

To find the least advantageous feasible weight vector $(v, u) \in (S_v, S_u)$ for DMU_o in terms of its efficiency score, we need to minimize its efficiency while ensuring that some $DMU_k \in \mathcal{D}$ is efficient. To guarantee that an efficiency score of some DMU is not less than one, we use binary variables b_k , $k = 1, \dots, K$. If $b_k = 1$, then $C(1 - 1) = 0$ and $\sum_{n=1}^N u_n y_{nk} \geq \sum_{m=1}^M v_m x_{mk}$; thus, $E_k(v, u) \geq 1$. Since we require that $\sum_{k=1}^K b_k \geq 1$, this condition needs to be satisfied for at least one DMU_k , $k = 1, \dots, K$. Otherwise, if $b_k = 0$, the use of C

prevents constraint violation. The minimization of $E_{o,*}$ in the objective function implies that a solver will assign ones to the binary variables so that to implement the least advantageous scenario for DMU_o .

2.2.2. Efficiency distribution

For each $DMU_o \in \mathcal{D}$, an efficiency acceptability interval index $EAI(DMU_o, b_i)$ is the share of feasible weight vectors $(v, u) \in (S_v, S_u)$ for which DMU_o attains an efficiency score in the interval $b_i \subset [0, 1]$ ($i = 1, \dots, B$, where B is the number of subintervals (buckets)). Let us denote with $b_{i,*}$ and b_i^* the extreme values of the subinterval b_i . Thus, $b_i = (b_{i,*}, b_i^*]$ with the proviso that b_1 is also left-closed (i.e., $b_1 = [b_{1,*} = 0, b_1^*]$). The buckets are constructed in the following way:

$$\bigcup_{i=1}^B b_i = [0, 1], \quad b_i \cap b_j = \emptyset, \quad i \neq j,$$

$$\text{and } b_i^* - b_{i,*} = b_{i+1}^* - b_{i+1,*}, \quad \text{for } i = 1, \dots, B-1.$$

While this is a default setting, in general, it is possible to construct buckets with different amplitudes so that $b_i^* - b_{i,*} \neq b_{i+1}^* - b_{i+1,*}$, for $i \in \{1, \dots, B-1\}$.

In the following we consider estimations EAI'_i of efficiency acceptability interval indices derived with Monte Carlo simulation. The same remark applies to pairwise efficiency outranking indices $PEOIs$ and efficiency rank acceptability indices $ERAI_s$ defined in Sections 2.3.2 and 2.4.2, respectively.

Proposition 2.1. For each $DMU_o \in \mathcal{D}$, $\sum_{i=1}^B EAI'_i(DMU_o, b_i) = 1$.

To enrich the view on the efficiency scores obtained in the representative sample $(S_v, S_u)^S$ of weight vectors (S_v, S_u) , we provide the following measures:

- the extreme efficiencies E_o^* and $E_{o,*}$ observed in $(S_v, S_u)^S \subset (S_v, S_u)$ for each $DMU_o \in \mathcal{D}$;
- an estimate of the expected efficiency $EE'_o = \sum_{(v,u) \in (S_v, S_u)^S} E_o(v, u)/W$, where W is the number of weight vectors in $(S_v, S_u)^S$.

2.3. Pairwise efficiency preference relations

In this section, we present the outcomes which materialize the outcomes of robustness analysis while referring to pairwise comparisons of DMUs. When compared to REA, we propose to consider a pair of efficiency preference relations instead of a single dominance relation. When compared to SMAA-D, we additionally analyze the pairwise efficiency outranking indices which indicate the probability that one DMU attains an efficiency at least as good as the other.

2.3.1. Possible and necessary efficiency preference relations

Applying all feasible weight vectors $(v, u) \in (S_v, S_u)$, we define two efficiency preference relations in the set of DMUs \mathcal{D} :

- Possible efficiency preference relation, \succeq_E^P , which is verified for a pair of DMUs $(DMU_o, DMU_k) \in \mathcal{D} \times \mathcal{D}$, in case $E_o(v, u) \geq E_k(v, u)$ holds for at least one $(v, u) \in (S_v, S_u)$;
- Necessary efficiency preference relation, \succeq_E^N , which is verified for a pair of DMUs $(DMU_o, DMU_k) \in \mathcal{D} \times \mathcal{D}$, in case $E_o(v, u) \geq E_k(v, u)$ holds for all $(v, u) \in (S_v, S_u)$.

The following LP needs to be considered to assess whether these relations hold:

$$\min/\max \quad E_o = \sum_{n=1}^N u_n y_{no}$$

$$\text{subject to : } \sum_{m=1}^M v_m x_{m0} = 1,$$

$$\sum_{n=1}^N u_n y_{nk} = \sum_{m=1}^M v_m x_{mk},$$

$$(v, u) \in (S_v, S_u). \tag{4}$$

If $E_o^{\max} = \max E_o$ obtained in problem (4) is not less than one, there exists some $(v, u) \in (S_v, S_u)$ for which $E_o(v, u) \geq E_k(v, u)$, and, thus, $DMU_o \succ_E^P DMU_k$. If $E_o^{\min} = \min E_o$ obtained in problem (4) is greater or equal to one, there is no feasible weight vector $(v, u) \in (S_v, S_u)$ for which $E_k(v, u) > E_o(v, u)$, and, thus, $DMU_o \succ_E^N DMU_k$.

In [53], the robustness analysis for pairs of DMUs is materialized with the efficiency dominance relation \succ_E . It holds for (DMU_o, DMU_k) if DMU_o necessarily attains the efficiency not less than DMU_k , while attaining strictly greater efficiency for some feasible weight vector. Thus, $DMU_o \succ_E DMU_k$ iff $DMU_o \succ_E^N DMU_k$ and $\neg(DMU_k \succ_E^N DMU_o)$. We consider a separate consideration of \succ_E^N and \succ_E^P (rather than aggregating these two results into \succ_E) more beneficial for the three following reasons:

- in case $DMU_o \succ_E DMU_k$, we may indicate whether DMU_k is possibly weakly preferred to DMU_o or not (i.e., whether $E_o(v, u) > E_k(v, u)$ for all $(v, u) \in (S_v, S_u)$, or for some $(v', u') \in (S_v, S_u)$, $E_o(v', u') = E_k(v', u')$);
- in case $\neg(DMU_o \succ_E DMU_k)$ and $\neg(DMU_k \succ_E DMU_o)$, we may indicate if DMU_o and DMU_k are related by the necessary indifference or necessary incomparability; in the former case, for all $(v, u) \in (S_v, S_u)$, $E_o(v, u) = E_k(v, u)$; in the latter case, for some $(v', u') \in (S_v, S_u)$, $E_o(v', u') > E_k(v', u')$ and for some $(v'', u'') \in (S_v, S_u)$, $E_k(v'', u'') > E_o(v'', u'')$;
- the possible and necessary efficiency preference relations converge with the growth of weight constraints provided by the DM (see Appendix C), while the dominance relation does not [53].

2.3.2. Pairwise efficiency outranking indices

For a pair of DMUs, $(DMU_o, DMU_k) \in \mathcal{D} \times \mathcal{D}$, a pairwise efficiency outranking index $PEOI(DMU_o, DMU_k)$ is the share of feasible weight vectors for which DMU_o is not worse than DMU_k in terms of the efficiency score, i.e., $E_o(v, u) \geq E_k(v, u)$.

Proposition 2.2. For $DMU_o \in \mathcal{D}$, $PEOI(DMU_o, DMU_o) = 1$.

Proposition 2.3. For $DMU_o, DMU_k \in \mathcal{D}$, $1 \leq PEOI(DMU_o, DMU_k) + PEOI(DMU_k, DMU_o) \leq 2$.

The pairwise efficiency winning index $PEWI(DMU_o, DMU_k)$ is the share of feasible weight vectors for which $E_o(v, u)$ is strictly better than $E_k(v, u)$.

Proposition 2.4. For $DMU_o, DMU_k \in \mathcal{D}$, $PEWI(DMU_o, DMU_k) = 1 - PEOI(DMU_k, DMU_o)$.

In the following we consider estimations of the pairwise efficiency indices $PEOI'$ and $PEWI'$ which are computed with Monte Carlo simulation.

2.4. Efficiency ranks

In this section, we discuss a set of results clearly indicating how the DMUs' efficiency ranks vary across the entire space of feasible weights. When compared to REA, to enhance understanding of the underlying logic, we discuss alternative formulations of linear programs for identifying the extreme ranks. When compared to SMAA-D, we propose to aggregate the rank acceptability indices into the estimates of expected efficiency rank for each DMU.

2.4.1. Extreme efficiency ranks

The rank of DMU_o relative to all DMUs in \mathcal{D} is defined with the ranking function:

$$R_o(v, u) = 1 + \sum_{DMU_k \in \mathcal{D} \setminus \{DMU_o\}} h(o, k, (v, u)), \quad \text{where} \tag{5}$$

$$h(o, k, (v, u)) = \begin{cases} 1, & \text{if } E_k(v, u) > E_o(v, u) \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

To identify the best $R_o^* = \min_{(v, u) \in (S_v, S_u)} R_o(v, u)$ efficiency rank that $DMU_o \in \mathcal{D}$ can attain, the following Mixed-Integer Linear Programming (MILP) model needs to be considered [53]:

$$\min R_o^* = 1 + \sum_{k=1, k \neq o}^K b_k$$

$$\text{subject to : } \sum_{n=1}^N u_n y_{no} = \sum_{m=1}^M v_m x_{m0} = 1,$$

$$[*] \sum_{n=1}^N u_n y_{nk} \leq \sum_{m=1}^M v_m x_{mk} + C b_k \quad (k = 1, \dots, K, k \neq o),$$

$$b_k \in \{0, 1\} \quad (k = 1, \dots, K, k \neq o),$$

$$(v, u) \in (S_v, S_u),$$

(7)

where C is a large positive constant. In the above problem, it is sufficient if:

$$C > \max_{DMU_o, DMU_k \in \mathcal{D}} \left\{ \max_{n=1, \dots, N} \{y_{nk}/y_{no}\} - \min_{m=1, \dots, M} \{x_{mk}/x_{m0}\} \right\}.$$

In problem (7), we identify the feasible weight vector $(v, u) \in (S_v, S_u)$ for which the number of DMUs with efficiency better than $E_o(v, u)$ is minimal. If $\sum_{n=1}^N u_n y_{nk}$ cannot be less or equal to $\sum_{m=1}^M v_m x_{mk}$ for some particular weight vector, a binary variable b_k corresponding to DMU_k , $k \neq o$, is instantiated with one. Then, being multiplied by a large positive constant C , $b_k=1$ prevents violation of constraint [*] for the respective k . This scenario occurs only if $E_k(v, u) = \sum_{n=1}^N u_n y_{nk} / \sum_{m=1}^M v_m x_{mk} > 1$, (i.e., if $\sum_{n=1}^N u_n y_{nk} - \sum_{m=1}^M v_m x_{mk} > 0$) while $E_o(v, u) = 1$. Then, $E_k(v, u) > E_o(v, u)$, and each $b_k=1$ identifies a unit ranked better than DMU_o . Otherwise, i.e., when $E_k(v, u) \leq E_o(v, u)$, b_k is instantiated with zero (then, $C b_k=0$). Since the objective function is minimized, the solver tries to assign as many zeros as possible to b_k , $k = 1, \dots, K, k \neq o$, thus, minimizing the cardinality of the set of DMUs which are ranked better than DMU_o . As a result, the sum of binary variables b_k , $k = 1, \dots, K, k \neq o$, increased by one is equal to the best (highest) rank of DMU_o . For example, in case there are three units simultaneously ranked better than DMU_o , $R_o^* = 3 + 1 = 4$.

The worst efficiency rank of DMU_o , $R_{o,*} = \max_{(v, u) \in (S_v, S_u)} R_o(v, u)$, is obtained as the optimum of the following MILP problem:

$$\max R_{o,*} = 1 + \sum_{k=1, k \neq o}^K b_k$$

$$\text{subject to : } \sum_{n=1}^N u_n y_{no} = \sum_{m=1}^M v_m x_{m0} = 1,$$

$$[*] \sum_{m=1}^M v_m x_{mk} \leq \sum_{n=1}^N u_n y_{nk} + C(1 - b_k) \quad (k = 1, \dots, K, k \neq o),$$

$$b_k \in \{0, 1\} \quad (k = 1, \dots, K, k \neq o), \quad (v, u) \in (S_v, S_u).$$

(8)

To prevent undesired compensations in the above problem, it is sufficient if:

$$C > \max_{DMU_o, DMU_k \in \mathcal{D}} \left\{ \max_{m=1, \dots, M} \{x_{mk}/x_{m0}\} - \min_{n=1, \dots, N} \{y_{nk}/y_{no}\} \right\}.$$

In problem (8), we identify the feasible weight vector $(v, u) \in (S_v, S_u)$ for which the number of DMUs with efficiency not worse than $E_o(v, u)$ is maximal. If $E_k(v, u) \geq E_o(v, u)$, a binary variable b_k is instantiated with one. Thus, the sum of binary variables b_k ,

$k = 1, \dots, K, k \neq o$, is equal to the number of DMUs simultaneously ranked not lower than DMU_o . When increased by one, this number indicates the worst rank of DMU_o .

To enhance understanding of the underlying reasoning, in Appendix A we present alternative formulations of the above MILPs.

2.4.2. Efficiency rank acceptability indices

For $DMU_o \in \mathcal{D}$ and rank $k = 1, \dots, K$, the efficiency rank acceptability index $ERAI(DMU_o, k) \in [0, 1]$, is the share of feasible weight vectors that grant DMU_o rank k .

Proposition 2.5. For each $DMU_o \in \mathcal{D}$, $\sum_{k=1}^K ERAI(DMU_o, k) = 1$.

In what follows, we consider Monte Carlo estimations of the efficiency rank acceptability indices $ERAI'$. They can be used to compute an estimate of the expected rank for $DMU_o \in \mathcal{D}$:

$$ER'_o = \sum_{k=1}^K k \cdot ERAI'(DMU_o, k).$$

In the Appendix, we provide additional relevant information concerning different types of discussed results. In Appendix B, we present the interdependencies between robust results and stochastic indices, thus, proving how they complement each other. In Appendix C, we elaborate on the evolution of results with incremental specification of weight constraints. Finally, in Appendix D, we discuss the impact of removing some DMUs from the considered set of units on the results.

3. Implementation on the Diviz platform

3.1. Diviz

Diviz is an open-source software which allows us to design, execute, and share complex workflows implementing procedures of decision analysis [42]. Even though it was originally designed for MCDA, its characteristics are general enough to account for methods of DEA. The software infrastructure consists of:

- a Java client for algorithmic workflow design and visual analysis of the outcomes,
- distant servers for executing the workflows, i.e., computing the results.

Decision analysis procedures as well as visualization or reporting tools are available in *diviz* via XMCDAs web-services. They need to read inputs and write outputs formatted using the XMCDAs standard. In this way, the web-services can interoperate and be combined into complex workflows.

3.2. Implemented methods for robustness analysis using ratio-based efficiency measure

Methods for robustness analysis using ratio-based data envelopment model have been implemented and made available on *diviz* as a

collection of individual components (modules). They can be subsequently used to construct complex algorithmic workflows. Each module requires three input files specifying, respectively, the list of DMUs, sets of inputs and outputs, and performance matrix. The linear weight constraints may be optionally provided in yet another input file. The modules implementing stochastic analysis need to be additionally provided with the number of weight vectors that should be sampled to compute the stochastic indices. The list of implemented modules is the following:

- *DEACCEfficiency* (computes E_o^* and $E_{o,*}$ for each $DMU_o \in \mathcal{D}$).
- *DEACCRPreferenceRelations* (verifies the truth of \succsim_E^P and \succsim_E^N for all pairs of DMUs).
- *DEACCExtremeRanks* (computes R_o^* and $R_{o,*}$ for each $DMU_o \in \mathcal{D}$).
- *DEASMAACCEfficiencies* (computes EAI_s' , E_o^* , $E_{o,*}$, and EE_o' for each for $DMU_o \in \mathcal{D}$; it requires specification of the number of efficiency subintervals (buckets) B and number of samples used in the Hit-And-Run algorithm).
- *DEASMAACCRPreferenceRelations* (computes $PEOI_s'$ for all pairs of DMUs; it requires specification of the number of samples), and
- *DEASMAACCRERanks* (computes $ERAI_s'$ for all DMUs and ranks; it requires specification of the number of samples).

To enrich the arsenal of methods that can be used to investigate efficiency of DMUs, we provide the following additional components: *CCRSuperEfficiency* (computes super-efficiency for each DMU [2]), *CCRCrossEfficiency* (computes cross-efficiency of each DMU either with an aggressive or benevolent approach [59,22]), and *CCREfficiency-Bounds* (computes four types of results: the minimal and maximal ratios of each DMU's efficiency and the best or the worst efficiency of any DMU [53]). Thanks to this, the practitioners can easily compare results of different methods, while teachers can present a wide spectrum of approaches to their students using the same data format and user interface. Moreover, all available DEA components in *diviz* are open-source, which enhances the addition of yet other methods by the researchers.

The structures of two exemplary modules, *DEACCEfficiency* and *DEASMAACCEfficiencies*, are presented in Figs. 1 and 2. They exhibit the required inputs, provided outputs, possible parametrization, and computation procedures.

3.3. Workflow design

The design of decision analysis workflows in *diviz* is performed via an intuitive graphical user interface. Each component is represented by a box which can be linked to data files or other computation modules. Thus, the design of the workflow does not require any programming skills, but rather understanding the role of each module [42]. To construct a workflow, the user chooses the modules (s)he is interested in from the list of available elements. Using a “drag-and-drop” function, (s)he adds them to the workspace along with the data files. Subsequently, the inputs and outputs of different components can be linked using connectors to define the structure

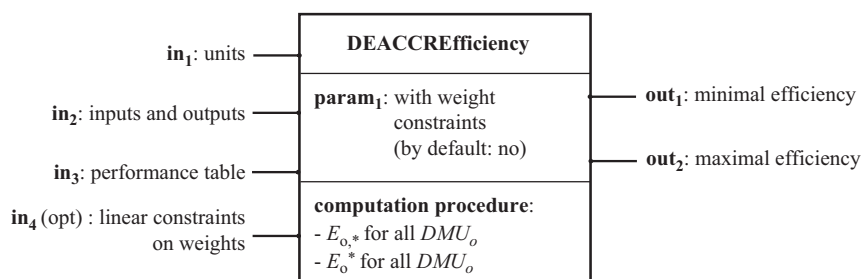


Fig. 1. Structure of *diviz* module which computes the extreme efficiency scores for each DMU using LP.

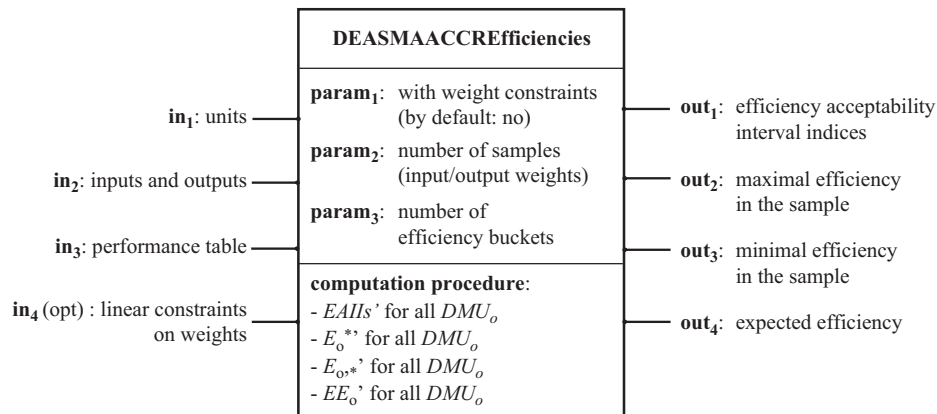


Fig. 2. Structure of *diviz* module which computes the efficiency acceptability interval indices, observed extreme efficiency scores, and expected efficiency for each DMU using Monte Carlo simulation.

of the workflow. In this way, the analysts may experiment with their own creations, suitably adjusting the arsenal of employed DEA methods to their own needs, while the researchers may design new software components that would built on the results delivered within our framework.

Once the design is finished, it is possible to execute the workflow. As already mentioned, the underlying calculations are performed on computing servers through the use of the XMCDAs web-services. Thus, *diviz* requires connection to the Internet. From the point of view of practitioners, this allows us to avoid performing heavy calculations on their local computers. The possibly multiple outcomes can be viewed either in *diviz* or in an external web-browser. The software maintains the history of all the past executions, which – in the context of efficiency analysis – is useful for studying the impact of additional weight constraints or removing the outlier DMUs on the results.

The *diviz* software enables to export any workflow as an archive (i.e., single file containing all necessary information including input data). This archive can be subsequently shared with other users, who can then import it (by loading the archive) into their software and execute it on the original data or continue the workflow's development. This is useful for the researchers for both dissemination and reproducibility of their results as well as for collaborative work on a particular case study.

Fig. 3 presents the workflow for our case study concerning analysis of efficiency of Polish airports, whose results are discussed in Section 4. Each module delivers different results based on the input data concerning the DMUs (*DEA-unit.xml*), definition of inputs and outputs (*DEA-inOut.xml*), and underlying performances (*DEA-performanceTable.xml*). Note that, e.g., to draw the graph of necessary efficiency preference relation, the appropriate output of the *DEACCRPreferenceRelations* module is provided as the input for *plotAlternativesHasseDiagram* module.

4. Application to efficiency analysis of Polish airports

4.1. Review of airport efficiency applications

As noted in [70], continuous improvement of airports' competitiveness greatly affects economic development of countries. Over the last twenty years DEA has proven its usefulness for studying the performance and measuring the efficiency of airports. Such examination is important from several points of view [70]. Firstly, governments or private owners can verify that the resources available to the airport are used as effectively as possible. Secondly, airlines and passengers want to use efficient airports. Thirdly,

managers can improve the competitiveness of the airports by following the best policy based on the competitors' performances.

The literature concerning DEA application to measuring the efficiency and productivity of the airports can be viewed from a few perspectives:

1. Employed model:

- CCR or BCC model for measuring airports' efficiency in a single year or season (e.g., [26,46,47,1,48,56,57,71]).
- DEA coupled with Malmquist productivity index to measure the airports' efficiency change over a few year period (e.g., [44,27,25,9]).
- DEA two-stage model, which first examines efficiency of the airports, and then uses a procedure to bootstrap DEA scores with a regression model for explanatory purpose (e.g., [8,7]).

2. Type of considered inputs:

- inputs related to the terminal services (e.g., number of check-in desks, gates, baggage collection belts, or parking spots, and terminal or baggage claim area) used, e.g., in [26,27,47,1,24,48,56,57,71,25,8];
- inputs related to the movement model (e.g., airport area, apron area, aircraft parking positions, numbers of runways and air routes connecting with other airports, runway length) used, e.g., in [26,27,48,72];
- monetary inputs (e.g., operational costs, labor costs, capital invested, capital stock, and airport charge) used, e.g., in [46,44,56,57,71,8];
- inputs related to the labor (e.g., number of employees) used, e.g., in [46,44,56,57];
- inputs related to the airport's localization (e.g., distance to the nearest city centre) used, e.g., in [1].

3. Type of considered outputs:

- outputs related to the terminal services (e.g., number of passengers, cargo throughput, and mail tonnes) used, e.g., in [26,46,44,27,47,48,56,71,25,8,72];
- outputs related to the movement model (e.g., aircraft movement, commuter movements, and number of air carrier operations) used, e.g., in [26,27,47,48,71,25,8,72];
- monetary outputs (e.g., total revenue, operational revenue, sales to plane, sales to passengers, commercial revenue, handling revenue, and non-aeronautical fee) used, e.g., in [56,57,8].

4. Geographical scope:

- single country (e.g., Argentina [8], Brazil [24], China [25], Italy [8], Japan [71], Spain [44,41], Turkey [35], United Kingdom [46,9], or United States [26,56,57]);
- continental or intercontinental scope (e.g., Europe [47,1,48] or Asia-Pacific region [65,70]).

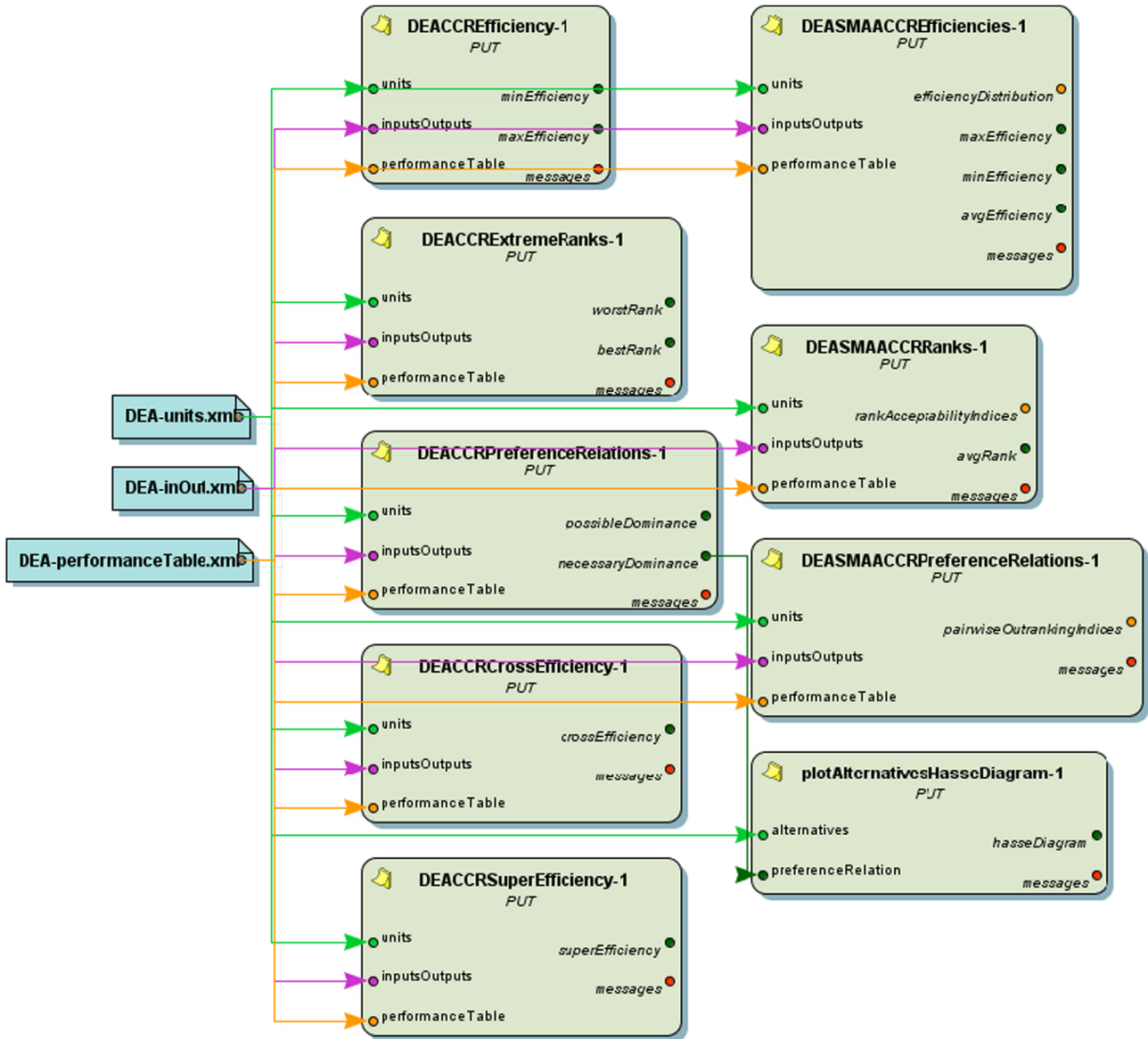


Fig. 3. Algorithmic workflow for the efficiency analysis of Polish airports.

4.2. Data description

We analyze data concerning performances of 11 Polish airports. The geographical distribution of the airports is presented in Fig. 4. Instead of the traditionally used basic inputs, such as the number of gates, aircraft parking positions, or runway length, we refer to more general and aggregated data on capacities of a terminal, runways, and an apron. These are derived from the report prepared by the world-wide leading consultancy companies [51] (see Table 1). As mentioned in [51], the values for $i_1 - i_3$ have been obtained directly from the airports. Additionally, we take into account a catchment area of each airport. The values for i_4 can be easily obtained from the Polish central statistical office. Detailed description of the four inputs is as follows:

i_1 : an annual capacity of a terminal defined as a passenger flow that an airport can accommodate without serious inconvenience (in million passengers per year); it takes into account



Fig. 4. Geographical distribution of Polish airports.

Table 1
Input and output performances for the problem of efficiency examination of Polish airports (all analyzed values concern 2009).

Airport	Short name	i_1	i_2	i_3	i_4	o_1	o_2
Warsaw	WAW	10.5	36	129.4	7.0	9.5	129.7
Cracow	KRK	3.1	19	31.6	7.9	2.9	31.3
Katowice	KAT	3.6	32	57.6	10.5	2.4	21.1
Wroclaw	WRO	1.5	12	18.0	3.0	1.5	18.8
Poznan	POZ	1.5	10	24.0	4.0	1.3	16.2
Lodz	LCJ	0.6	12	24.0	3.9	0.3	4.2
Gdansk	GDN	1.0	15	42.9	2.5	2.0	23.6
Szczecin	SZZ	0.7	10	25.7	1.9	0.3	4.2
Bydgoszcz	BZG	0.3	6	3.4	1.2	0.3	4.2
Rzeszow	RZE	0.6	6	11.3	2.7	0.3	3.5
Zielona Gora	IEG	0.1	10	63.4	3.0	0.005	0.61

limits on the traffic related to the terminal area, the numbers of gates and check-in counters, as well as severe congestion in access facilities;

- i_2 : a maximal throughput capacity defined as an average number of movements (arrivals and/or departures) that can be performed on the airport's runways (in number of movements per hour); it accounts for the configuration of runways, taxiways, waiting areas, and high speed exits, air traffic flow in the runway area (including an average runway occupancy time), and air traffic delays to the landing and takeoff moments;
- i_3 : a dynamic apron capacity defined as an average number of planes that can be served by the airport (in number of planes per hour); it is derived from the number and configuration of stands and ramps as well as an average stand occupancy time;
- i_4 : a catchment area of an airport defined as the number of inhabitants living within the range of 100 km from the airport (in million inhabitants); it reflects the airport's potential for attracting the surrounding population.

When it comes to the outputs, we focus on the two primary indicators related to the terminal services and movement model, defined in the following way:

- o_1 : passengers traffic measured by the total number of passengers served by the airport (in million passengers per year);
- o_2 : number of aircraft movements (one total movement is a landing or takeoff of an aircraft) (in thousand movements per year).

The outputs are derived from the statistical data provided by the Civil Aviation Authority (CAA) in Poland [16] (see Table 1, columns $o_1 - o_2$). Let us emphasize that we have carefully selected the inputs and outputs so that they harmonize. Indeed, the inputs of each airport reflect its individually judged potential, whereas the outputs indicate the degree to which this potential is used in practice.

4.3. Results

In this section, we discuss results derived from robustness and stochastic analysis of efficiency of Polish airports. As proven in the review presented in Section 4.1, such comprehensive analysis has never been conducted for any airport efficiency application. Moreover, while there exist some reports on measuring the efficiency of airports in many other countries, this aims to be the first comprehensive study for Poland.

First, we focus on the efficiency scores; then, we elaborate on the efficiency ranks; we conclude with the efficiency preference relations. All stochastic results presented in this section were derived from the analysis of 10 000 input and output weights

obtained with a Hit-And-Run algorithm. Then, we illustrate the impact of considering weight constraints and eliminating some outlier DMU.

For this purpose, we have constructed dedicated *diviz* workflows which are available online¹:

- *DEAPolishAirports.dvz* for results presented in Sections 4.3.1, 4.3.2 and 4.3.3 without considering weight constraints.
- *DEAPolishAirportsWithConstraints.dvz* for results discussed in Section 4.3.4 when considering weight constraints.
- *DEAPolishAirportsWithoutOutlier.dvz* for results discussed in Section 4.3.5 when considering the set of airports without WAW.

These workflows can be used to reproduce the results discussed in this section. For this purpose: (1) download *diviz*,² (2) launch it, (3) import the workflow ("Workflow - Import as new"), (4) run it on *diviz* ("Execution - Run"), and (5) view the results of interest by selecting a particular module's output. Moreover, they illustrate how to prepare the input data so that they can be later easily adapted to other problems.

4.3.1. Efficiency distribution and extreme efficiencies

Table 2 (columns E_o^* and $E_{o,*}$) shows the best and the worst efficiency scores for each DMU, $DMU_o \in \mathcal{D}$. Five airports with $E_o^* = 1$ (WAW, KRK, WRO, GDN, and BZG) are deemed as efficient. Among the six inefficient airports with $E_o^* < 1$, POZ and IEG have, respectively, the least and the greatest gap that needs to be covered for reaching efficiency. Their maximal efficiency scores are equal to 0.799 and 0.258, respectively. The minimum efficiencies $E_{o,*}$ for all airports are less than 0.5. This means that for the least advantageous weight vector for each DMU, it is at least twice less efficient than another DMU. Interestingly, when taking into account the worst efficiency scores, POZ (judged inefficient) compares positively to KRK and BZG (judged efficient).

The efficiency acceptability interval indices are provided in Table 3. We used 10 efficiency buckets with the same amplitude of 0.1. While for some airports the vast majority of attained efficiency scores is concentrated within a single bucket (e.g., for WAW in (0.9, 1.0], or LCJ and SZZ in (0.1, 0.2]), for some other airports the distribution of scores is more balanced. In particular, for BZG the probability of attaining efficiency in seven different ranges between (0.3, 0.4] and (0.9, 1.0] is greater than 8%. Analogously, for WRO three out of ten different *EAIIs'* are greater than 20%.

It is worthwhile analyzing the *EAIIs'* along with the extreme efficiencies observed in the sample (see columns E_o^* and $E'_{o,*}$ in Table 2). For most airports these differ from the true extreme efficiency scores computed with LP. In particular, for RZE, $E_{RZE}^* = 0.409 > E'_{RZE} = 0.359$ and $E_{RZE,*} = 0.069 < E'_{RZE,*} = 0.085$, whereas for IEG, $E_{IEG}^* = 0.258 > E'_{IEG} = 0.051$. Such analysis allows us to identify the ranges of scores which are attained only for marginal share of feasible weight vectors. In this perspective, the estimates of *EAIIs* derived from Monte Carlo simulation may be equal to 0.0, while there exists some feasible input/output weight vector (not included in the sample) for which a DMU would attain efficiency contained in the underlying bucket (see, e.g., *EAII*(WAW, (0.4, 0.5]) or *EAII*(IEG, (0.1, 0.2])).

Finally, the estimates of expected efficiency EE'_o (see column EE'_o in Table 2) may be used to rank the airports. In this case, WAW significantly outperforms other cities with $EE'_{WAW} = 0.944$, and IEG is placed at the very bottom with $EE'_{IEG} = 0.010$. When compared to cross-efficiencies (see column CE_o in Table 2), the advantage of

¹ <http://www.decision-deck.org/diviz/workflows.html>

² <http://www.decision-deck.org/diviz/download.html>

(SZZ) is ranked between 2 and 5 (9 and 10), whereas, in general, its rank interval is [1, 6] ([7, 10]). Finally, six airports have some rank acceptability indices equal to 0.0, even though analysis of the exact extreme results indicates that they may be possibly attained for at least one weight vector ((WAW, 5), (KAT, 9–10), (LCJ, 7–8), (BZG, 8), (RZE, 8, 10–11), (IEG, 8–10)). For each airport, ERAIs' can be aggregated into the estimates of an expected rank (see Table 4, column ER'_o). The airports with low ER'_o 's (e.g., WAW, WRO, and KRK) are average good performers, while the units with high ER'_o 's (e.g., RZE, SZZ, LCJ, and IEG) are on average far from being efficient, being ranked lower than the majority of airports.

4.3.3. Pairwise efficiency outranking indices and necessary/possible efficiency preference relations

The necessary and possible preference relations are provided in Table 5. Obviously, the truth of necessary efficiency relation implies the truth of a less demanding possible relation (for clarity, in Table 5 we list only these possible relations which are not necessary at the same time). There are 32 pairs of airports (DMU_o, DMU_k) $\in \mathcal{D} \times \mathcal{D}$, $o \neq k$, related by the necessary preference.

Table 5
Necessary and possible efficiency preference relations.

Airport	Necessary preference	Additional possible preference
WAW	KAT, POZ, LCJ, SZZ, RZE, IEG	WAW KRK, WRO, GDN, BZG
KRK	KAT, LCJ, SZZ, RZE, IEG	KRK WAW, WRO, POZ, GDN, BZG
KAT	RZE	KAT LCJ, SZZ, BZG, IEG
WRO	KAT, LCJ, SZZ, RZE, IEG	WRO WAW, KRK, POZ, GDN, BZG
POZ	KAT, LCJ, SZZ, RZE, IEG	POZ KRK, WRO, GDN, BZG
LCJ	IEG	LCJ KAT, SZZ, RZE
GDN	KAT, LCJ, SZZ, RZE, IEG	GDN WAW, KRK, WRO, POZ, BZG
SZZ	KAT, LCJ, SZZ, RZE, IEG	SZZ KAT, LCJ, RZE, IEG
BZG	LCJ, SZZ, RZE, IEG	BZG WAW, KRK, KAT, WRO, POZ, GDN
RZE		RZE LCJ, SZZ, BZG, IEG
IEG		IEG KAT, SZZ, RZE

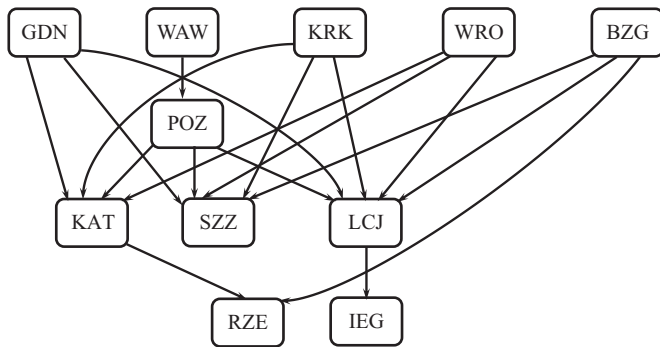


Fig. 5. The necessary efficiency preference relation.

Table 6
Pairwise efficiency outranking indices (in %).

Airport	WAW	KRK	KAT	WRO	POZ	LCJ	GDN	SZZ	BZG	RZE	IEG
WAW	100.0	98.01	100.0	95.41	100.0	100.0	99.84	100.0	71.37	100.0	100.0
KRK	1.99	100.0	100.0	18.35	99.56	100.0	83.45	100.0	43.53	100.0	100.0
KAT	0.00	0.00	100.0	0.00	0.00	100.0	0.00	99.97	0.56	100.0	100.0
WRO	4.59	81.65	100.0	100.0	99.58	100.0	89.13	100.0	53.25	100.0	100.0
POZ	0.00	0.44	100.0	0.42	100.0	100.0	62.51	100.0	26.86	100.0	100.0
LCJ	0.00	0.00	0.00	0.00	0.00	100.0	0.00	22.29	0.00	0.00	100.0
GDN	0.16	16.55	100.0	10.87	37.49	100.0	100.0	100.0	31.53	100.0	100.0
SZZ	0.00	0.00	0.03	0.00	0.00	77.71	0.00	100.0	0.00	2.82	100.0
BZG	28.63	56.47	99.44	46.75	73.14	100.0	68.47	100.0	100.0	100.0	100.0
RZE	0.00	0.00	0.00	0.00	0.00	100.00	0.00	97.18	0.00	100.0	100.0
IEG	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.0

For example, WAW and BZG are necessarily preferred to, respectively, six and four other airports.

The graph of necessary efficiency relation, subject to a transitive reduction, is illustrated in Fig. 5. When analyzing this graph, the efficient airports are confirmed to be the best DMUs, because there is no other airport which is necessarily preferred to them. Note, however, that the inverse implication is not true, i.e., there may exist some inefficient DMU such that there is no other DMU necessarily preferred to it. Among the inefficient airports, POZ confirms its necessary superiority over the five remaining inefficient DMUs. Further, IEG, RZE, and SZZ should be viewed as the worst airports, because they are not necessarily preferred to any other airports.

The analysis of the diagram may be enriched with the view on the possible relations. For example, on one hand, LCJ is not possibly preferred to BZG, which means that the efficiency of BZG is always strictly greater than that of LCJ. On the other hand, although BZG is necessarily preferred to RZE, the latter is possibly preferred to the former. This means that there is at least one weight vector for which efficiencies of these two airports are equal.

Furthermore, it is interesting to analyze the graph of the necessary relation in the context of extreme ranks. Some of the observed interdependencies are straightforward. For example, while GDN (LCJ) is necessarily preferred to five (by six) other airports, its worst (best) rank is $11 - 5 = 6$ ($1 + 6 = 7$). However, some other results are not that obvious. For example, POZ is necessarily preferred only by WAW, but its best rank is 3, whereas SZZ is not necessarily preferred to any other airport, but it is not ranked at the very bottom in the worst case.

Finally, let us note that the nodes which are not related by an arc in the diagram, indicate the airports which are incomparable in terms of \succ_E^N (e.g., (WAW,BZG), (POZ,WRO), (BZG,KAT), or (LCJ, RZE)). For such pairs, one of the airports is possibly (for some weight vector) more efficient than the other, and vice versa. When considering the outcomes of the traditional robustness analysis, these pairs are left incomparable (no additional information is given). Instead of leaving the analyst only with information that the possible preference relations are observed for at least one weight vector, our approach provides estimates of the shares of weight vectors confirming these outcomes.

In Table 6, we present pairwise efficiency outranking indices for all DMUs. Obviously, for pairs of airports related by the necessary relation (e.g., (WAW,KAT) or (GDN,RZE)), the respective $PEOI'$ is 100%, while for pairs not related by the possible relation (e.g., (KAT,KRK) or (SZZ,WRO)), $PEOI'$ is 0%. When it comes to pairs related by the necessary incomparability, for some of them one airport is more efficient than the other for the vast majority of weight vectors. In particular, for (WAW,KRK) $PEOI'(WAW, KRK) = 98.01\%$ and $PEOI'(KRK, WAW) = 1.99\%$, and for (WRO,POZ), $PEOI'(WRO, POZ) = 99.58\%$ and $PEOI'(POZ, WRO) = 0.42\%$. As for the

efficient airports, analysis of the pairwise efficiency outranking indices supports WAW in comparison with KRK, WRO, GDN, and BZG. For some pairs of airports, indicating the more advantageous one on the basis of *PEOIs* is not possible. For example, for (WRO, BZG), $PEOI'(WRO, BZG) = 53.25\%$ and $PEOI'(BZG, WRO) = 46.75\%$. Finally, although $PEOI'$ for (RZE, BZG) is equal to 0%, the possible efficiency relation for this pair holds, whereas even though $PEOI'$ for (SZZ, IEG) is equal to 100%, the necessary preference relation for this pair does not hold.

As justified in Section 2.1, when conducting Monte Carlo simulation, we assumed a uniform weight distribution for the space of feasible weights. Let us emphasize that with other exogenously given distribution the values of efficiency acceptability indices could be different. This is partially illustrated in Section 4.3.4, when a value of a density function assigned to some weight vectors is zeroed, because they are excluded from the feasible space by the provided weight constraints.

4.3.4. Incremental specification of weight constraints

For illustrative purpose, in this section, we assume that the following set of linear weights constraints has been provided by the DM:

- input weights: $v_1 \geq 3v_3$, $v_1 \geq 5v_3$, $v_2 \geq 2v_3$, and $v_2 \geq 5v_4$;
- output weights: $u_1 \geq 5u_2$.

In Table 7, we provide extreme efficiency scores and ranks in two iterations, i.e., when considering the weight space without (1) and with (2) the above specified constraints. These illustrate that the ranges of efficiencies and ranks become more precise when preference information is taken into account. In particular, KRK, WRO, and BZG become not efficient. Their best efficiency score is less than one ($E_{0,2}^* < 1$ and $E_{0,1}^* = 1$) and they are ranked second in the best case ($R_{0,2}^* = 2$ and $R_{0,1}^* = 1$). This implies that only WAW and GDN remain efficient. Constraining the weight space is neither advantageous for IEG. Its best efficiency score drops from 0.258 to 0.188, while the best rank decreases from 8 to 11. As a result, IEG is ranked at the very bottom for all feasible weight vectors. Furthermore, with limited weight space, GDN attains the best lowest efficiency score ($E_{GDN,*2} = 0.455 > E_{WAW,*2} = 0.452$), while for KRK and POZ the increase of the worst efficiency is greater than 0.2. Even though their lowest scores are much better now, their least ranks remain unchanged. On the contrary, RZE (KAT) is now ranked 9 (8) for the least advantageous weight vector, while it was ranked 11 (10) without weight constraints.

In Fig. 6, we depict the graph of the necessary efficiency preference relation derived from the analysis of constrained weight space. This graph is enriched when compared with the one presented in Fig. 5. Precisely, there are five pairs for which the necessary relation has become true: (KAT, RZE), (KAT, IEG), (SZZ, IEG), (RZE, SZZ), and (RZE, IEG). Interestingly, even though KRK, WRO, and BZG are not efficient and ranked second in the best case, there is no other airport that would be necessarily preferred to them. This confirms the benefits of joint consideration of the three outcome perspectives: scores, ranks, and preference relations.

To illustrate the effect of incorporating weight constraints on the acceptability indices, in Table 8 we present *EALs'*, *ERAls'*, and *PEOIs'* for BZG without and with weight constraints. The most evident effect of integrating these constraints into the efficiency model is that for the vast majority of feasible weight vectors (about 97%) BZG attains efficiency scores in the range (0.2, 0.6] and ranks 6–7, while previously it attained the best efficiency scores (0.9, 1.0) and ranks on the podium (1–3) for, respectively, over 35% and 50%. Moreover, BZG is now far less advantageous when compared with WAW, KRK, WRO, POZ, and GDN.

Table 7
Extreme efficiency scores and ranks without (1) and with (2) weight constraints.

Airport	$E_{0,1}^*$	$E_{0,*1}$	$E_{0,2}^*$	$E_{0,*2}$	$R_{0,1}^*$	$R_{0,*1}$	$R_{0,2}^*$	$R_{0,*2}$
WAW	1.000	0.452	1.000	0.452	1	5	1	5
KRK	1.000	0.213	0.962	0.439	1	6	2	6
KAT	0.591	0.108	0.554	0.210	6	10	6	8
WRO	1.000	0.338	0.922	0.445	1	5	2	5
POZ	0.799	0.218	0.779	0.433	3	6	3	6
LCJ	0.300	0.057	0.282	0.094	7	10	7	10
GDN	1.000	0.302	1.000	0.455	1	6	1	6
SZZ	0.271	0.089	0.260	0.113	7	10	9	10
BZG	1.000	0.184	0.954	0.189	1	8	2	8
RZE	0.409	0.069	0.383	0.169	7	11	7	9
IEG	0.258	0.001	0.188	0.001	8	11	11	11

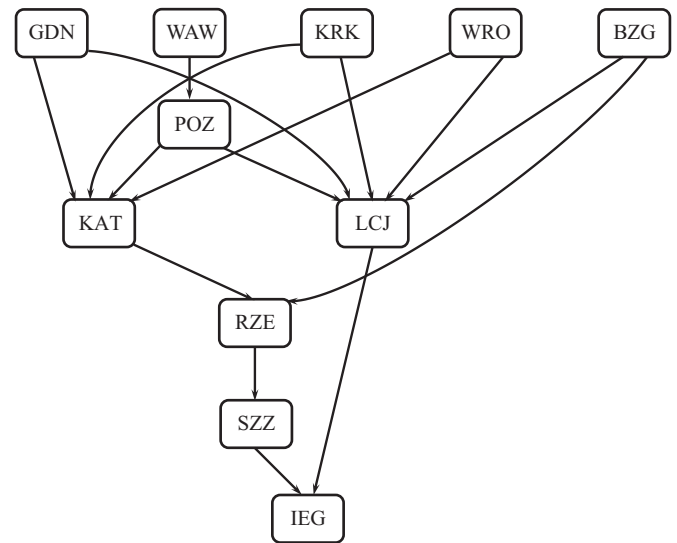


Fig. 6. The necessary preference relation when accounting for the weight constraints.

4.3.5. Elimination of outlier DMUs

For illustrative purpose, in this section, we investigate the impact of removing some outlier airports on the obtained results. We refer to the backward approach presented in [6], and eliminate the units with super-efficiency greater than 2.0. Subsequently, we compare the original results from Sections 4.3.1–4.3.3 with the ones obtained while neglecting WAW.³ Note that WAW is the largest and busiest airport in Poland, which also proved to be the best in terms of robustness analysis in our results.

In Table 9, we provide the extreme efficiency scores and ranks as well as the estimates of expected efficiencies and efficiency ranks for the set of airports with (D_1) and without (D_2) considering WAW. For all airports, the extreme efficiency scores are not worse when WAW is neglected. Precisely, for POZ and SZZ the best efficiencies have been improved. In fact, POZ is the greatest beneficiary of removing WAW from the analysis ($E_{POZ}^{*,D_2} = 0.989 > E_{POZ}^{*,D_1} = 0.799$). Furthermore, the worst efficiencies have been improved significantly for all airports (e.g., $E_{WRO}^{*,D_2} = 0.5 > E_{WRO}^{*,D_1} = 0.338$) but GDN, SZZ, and IEG. When it comes to the extreme ranks, all airports have improved their worst ranks by one (e.g., $R_{BZG,*}^{D_2} = 7 < R_{BZG,*}^{D_1} = 8$). The same holds for the non-efficient airports in terms of their best ranks (e.g., $R_{POZ}^{*,D_2} = 2 < R_{POZ}^{*,D_1} = 3$). This

³ In diviz, elimination of some DMU from the analysis can be conducted easily by setting a unit-specific attribute “active” to “false”.

Table 8

Efficiency acceptability interval indices (in %), rank efficiency acceptability indices (in %), and pairwise efficiency outranking indices (in %) for BZG without (1) and with (2) weight constraints.

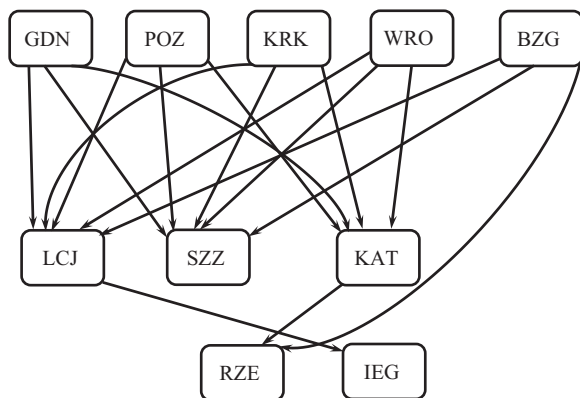
Indices	[0.0, 0.1]	(0.1, 0.2]	(0.2, 0.3]	(0.3, 0.4]	(0.4, 0.5]	(0.5, 0.6]	(0.6, 0.7]	(0.7, 0.8]	(0.8, 0.9]	(0.9, 1.0]
EAI'_1	0.00	0.04	1.59	11.35	11.10	11.73	10.35	8.93	8.78	36.16
EAI'_2	0.00	0.00	16.04	40.99	26.75	12.75	2.97	0.40	0.10	0.00
	1	2	3	4	5	6	7	8	9	10–11
$ERAI'_1$	28.56	17.49	7.00	17.81	4.58	23.98	0.58	0.00	0.00	0.00
$ERAI'_2$	0.00	0.01	0.07	0.40	1.22	84.45	13.85	0.00	0.00	0.00
	WAW	KRK	KAT	WRO	POZ	LCJ	GDN	SZZ	RZE	IEG
$PEOI'_1$	28.63	56.47	99.44	46.75	73.14	100.0	68.47	100.0	100.0	100.0
$PEOI'_2$	0.02	0.20	85.65	0.06	1.25	100.0	0.06	100.0	100.0	100.0

Table 9

Extreme efficiency scores and ranks, estimates of expected efficiencies and efficiency ranks with (D_1) and without (D_2) considering WAW.

Airport	E_0^{*,D_1}	$E_{0,*}^{D_1}$	E_0^{*,D_2}	$E_{0,*}^{D_2}$	$EE_0^{D_1}$	$EE_0^{D_2}$	R_0^{*,D_1}	$R_{0,*}^{D_1}$	R_0^{*,D_2}	$R_{0,*}^{D_2}$	$ER_0^{D_1}$	$ER_0^{D_2}$
WAW	1.000	0.452	–	–	0.944	–	1	5	–	–	1.3534	–
KRK	1.000	0.213	1.000	0.420	0.664	0.854	1	6	1	5	3.5354	2.5769
KAT	0.591	0.108	0.591	0.220	0.281	0.362	6	10	5	9	6.9947	5.9953
WRO	1.000	0.338	1.000	0.500	0.702	0.901	1	5	1	4	2.7192	1.7638
POZ	0.799	0.218	0.989	0.370	0.533	0.699	3	6	2	5	5.0994	4.1006
LCJ	0.300	0.057	0.300	0.095	0.133	0.174	7	10	6	9	9.7795	8.7795
GDN	1.000	0.302	1.000	0.302	0.531	0.707	1	6	1	5	5.0322	4.0201
SZZ	0.271	0.089	0.274	0.089	0.145	0.192	7	10	6	9	9.1935	8.1943
BZG	1.000	0.184	1.000	0.312	0.726	0.891	1	8	1	7	3.2662	2.5440
RZE	0.409	0.069	0.409	0.137	0.221	0.286	7	11	6	10	8.0265	7.0305
IEG	0.258	0.001	0.258	0.001	0.010	0.014	8	11	7	10	11.000	10.000

a



b

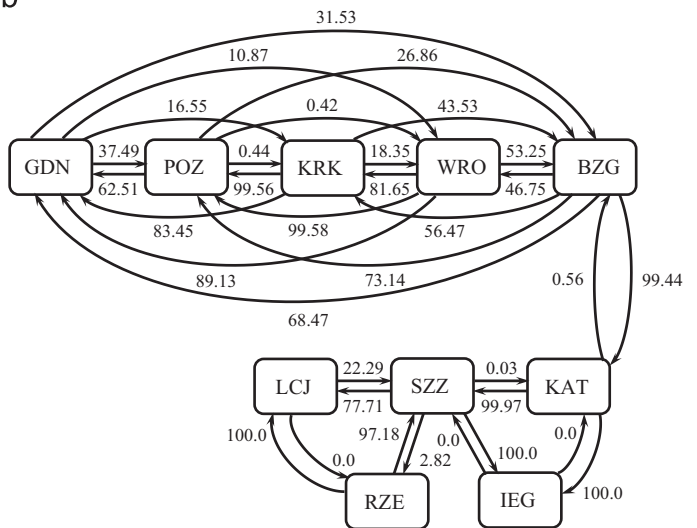


Fig. 7. The necessary preference relation (a) and pairwise efficiency outranking indices (b) without considering WAW.

confirms the robustness of the ranking intervals, which change at most by one when a single unit is removed or introduced.

As far as the expected efficiency scores and ranks are concerned, the estimates of these measures obtained from the Monte Carlo simulation after removing WAW indicate a clear improvement (i.e., greater expected efficiency and less expected rank) for all airports (e.g., $EE_{KRK}^{D_2} = 0.854 > EE_{KRK}^{D_1} = 0.664$ and $ER_{KRK}^{D_2} = 2.5769 < ER_{KRK}^{D_1} = 3.5354$).

In Fig. 7, we depict the graph of the necessary efficiency preference relation and pairwise efficiency outranking indices derived from the analysis neglecting WAW. For all pairs of airports, the

truth or falsity of \approx_E^N as well as the values of $PEOIs$ are the same as in Fig. 5 and Table 6, respectively.⁴ This confirms that the removal of some outlier DMUs does not influence the pairwise one-on-one results for the remaining units.

Finally, let us remind that the convergence of results with the removal/introduction of some unit cannot be predicted in case of

⁴ In general, the estimates of the pairwise efficiency outranking indices may differ slightly from one Monte Carlo simulation to another because there is no guarantee that the sets of feasible weight vectors sampled in these simulations are the same.

Table 10
Efficiency acceptability interval indices (in %) and rank efficiency acceptability indices (in %) for BZG with (\mathcal{D}_1) and without (\mathcal{D}_2) considering WAW.

Indices	[0.0, 0.1]	(0.1, 0.2]	(0.2, 0.3]	(0.3, 0.4]	(0.4, 0.5]	(0.5, 0.6]	(0.6, 0.7]	(0.7, 0.8]	(0.8, 0.9]	(0.9, 1.0]
$EAII^{\mathcal{D}_1}$	0.00	0.04	1.59	11.35	11.10	11.73	10.35	8.93	8.78	36.16
$EAII^{\mathcal{D}_2}$	0.00	0.00	0.00	0.11	0.54	2.58	9.08	13.90	14.53	59.26
	1	2	3	4	5	6	7	8	9	10
$ERAI^{\mathcal{D}_1}$	28.56	17.49	7.00	17.81	4.58	23.98	0.58	0.00	0.00	0.00
$ERAI^{\mathcal{D}_2}$	46.38	7.25	17.23	4.41	24.19	0.54	0.00	0.00	0.00	0.00

efficiency acceptability interval indices and efficiency rank acceptability indices. To illustrate this phenomenon, in Table 10, we present the $EAIIs'$ and $ERAIIs'$ for BZG obtained with and without considering WAW. The share of compatible weight vectors for which BZG attains the most advantageous efficiency scores (0.9, 1.0) and rank (1) has now increased, but for the remaining stochastic results one cannot observe any regularities except the ones already captured with the expected efficiency scores and ranks.

5. Practical consideration

5.1. In what contexts is the proposed framework relevant?

The proposed integrated framework for robustness analysis should be used when facing at least one of the following characteristics:

1. The management wishes to investigate the performance of DMUs for all feasible input/output weights and/or their significant share. It is desirable, because the feasible weight vectors reflect relevant preference information and represent the full spectrum of priorities that can be assigned to different inputs and outputs. This allows us to judge the goodness of DMUs as overall performers in a more reliable way than in the traditional DEA approaches, which take into account only extremely small share of feasible weights.
2. The analyst does not want or know how to formulate, or is not able to justify the assumptions about possible returns to scale. In the proposed framework, the production possibilities are defined only by the considered DMUs, and the results are derived from pairwise comparisons among the existing units rather than measuring efficiency relative to the efficient frontier. This makes the results more reliable and less sensitive to changing the set of DMUs.
3. The number of compared DMUs is relatively small. Indeed, our framework can be used with any number of DMUs, even if it is not large enough compared to the overall number of inputs and outputs as required for the interpretability of traditional DEA outcomes (e.g., maximal efficiency scores, cross-efficiencies, or super-efficiencies). In our case study, the number of analyzed airports (11) is not significantly greater than the number of inputs and outputs (6). Nonetheless, the proposed approach can still provide interpretable and valuable results derived from the one-on-one comparisons of airports.
4. The management is interested in an in-depth analysis that would concern at least one of the following perspectives on DMUs' efficiency: scores, ranks, or preference relation. Firstly, the scores determine how much worse is a given DMU than the most efficient unit. Secondly, the ranks indicate how many DMUs are better/worse than a given DMU in terms of their efficiency ratio. Thirdly, the preference relation offers a unique one-on-one perspective for the efficiency analysis instead of one-against-all viewpoint being more typical for DEA.

These three perspectives are complementary, and the results offered by one of them, in general, cannot be derived from the analysis of another. Let us provide some examples supporting this claim:

- if there is no unit that would be necessarily preferred to a given DMU, it may still be not efficient, thus, attaining an efficiency score less than one in the best case (see, e.g., the case of POZ in Section 4.3.5 which is not necessarily preferred by any other airport, but attains $E_{POZ}^{*,\mathcal{D}_2} = 0.989 < 1$);
- if the intersection of the ranges of possible efficiency ranks for a pair of DMUs is non-empty, one of them may be still necessarily preferred to another (see, e.g., the case of KAT and RZE in Section 4.3.5 with $KAT \succ_{E}^{N,\mathcal{D}_2} RZE$ and $[R_{KAT}^{*,\mathcal{D}_2}, R_{*,KAT}^{\mathcal{D}_2}] \cap [R_{RZE}^{*,\mathcal{D}_2}, R_{*,RZE}^{\mathcal{D}_2}] = [6, 9]$); furthermore, the pairwise outranking index may indicate that the vast majority of feasible weights ranks higher a DMU with less advantageous ranking interval (see, e.g., the case of SZZ and RZE in Sections 4.3.1–4.3.3, where $R_{SZZ}^* \leq R_{RZE}^*$ and $R_{*,SZZ} \leq R_{*,RZE}$, but $PEOI'(RZE, SZZ) = 97.18\%$);
- if the least efficiency indicates a significantly worse performance of a given DMU when compared to the most efficient unit, it can be still better than the vast majority of other DMUs as proven by its worst possible rank (see, e.g., the case of GDN in Section 4.3.5 with $E_{GDN,*}^{\mathcal{D}_2} = 0.302$, thus, being over 3 times less efficient than the most efficient airport, while still being ranked better than 5 out of 9 other airports ($R_{GDN,*}^{\mathcal{D}_2} = 5$)).

Nonetheless, in practical efficiency analysis, one can use only a small subset of results that may be delivered within the proposed framework.

5.2. How to interpret different robust results and which managerial concerns they address?

Traditionally, DEA has been used for indicating which DMUs are efficient and inefficient, thus discriminating only between these two groups. In some real-world situations, the shares of efficient or inefficient DMUs may be very large, and the management may wish to identify a small subset of the most distinguishing ones among them. In their work, Tsou and Huang [64] discuss several ranking methods that have been proposed to improve the discrimination power of DEA. Our framework derives from the fact that each feasible weight vector provides a basis for the performance comparison, thus, offering greater discrimination among DMUs. The robust results which synthesize the outcomes obtained for different weight vectors can be used for answering the following relevant questions (we will provide the exemplary answers to these questions while referring to the results of our case study presented in Sections 4.3.1–4.3.3):

1. *Which efficient DMUs perform well compared to other DMUs?* For the efficient DMUs with $E_o^* = 1$ and $R_o^* = 1$, one should consider how frequently they attain the best ranks and efficiency scores and how bad they can be at worst (i.e., $E_{*,o}$ and $R_{*,o}$). This allows us

to distinguish the overall good performers exhibiting more universal good practices to follow from the more niche DMUs which are efficient only under very specific conditions, while being far from efficiency for the vast majority of feasible weights. In our study, the best examples of the former group are WAW and BZG, while GDN is the most representative for the latter group. Following these conclusions, WAW and BZG are advised to be used for benchmarking. Such discrimination between the efficient DMUs may also stimulate data mining to generate hypotheses about the drivers of strong or weak efficiency.

2. *Which inefficient DMUs do not perform significantly worse compared to other DMUs?* For the inefficient DMUs with $E_o^* < 1$ and $R_o^* < 1$, one should analyze how good they can be at best (i.e., E_o^* and R_o^*) as well as how often they attain their worst ranks and efficiency scores. This allows us to discriminate between the inefficient DMUs which have the greatest potential for becoming efficient and these for which attaining efficiency would be most challenging. The management may decide to implement the corrective actions for the former group in the first order, while the latter group seems to be crucial in terms of reducing the performance gap between the best and worst performers. In our study, the most advantageous inefficient airport is POZ, while KAT, RZE, SZZ, LCJ, and IEG require considerable improvement in their efficiency ratios, which is confirmed by a large spectrum of priorities that can be assigned to the inputs and outputs.
3. *Which DMUs are the average good or bad performers?* Irrespective of their efficiency status, all DMUs can be ranked from the best to the worst based on their average efficiency scores (EE_o) or ranks (ER_o). These results compare the DMUs using a large number of weight vectors, clearly exhibiting which units perform good for different priorities that can be assigned to inputs and outputs. In some situations, the expected efficiency scores or ranks can prove that inefficient DMUs are on average better than some efficient ones, thus, indicating the need for possible corrective actions also in the context of the efficient units. In our study, this is the case of GDN, since the comparison of expected efficiencies of POZ (deemed inefficient) and GDN (judged efficient) indicates that $EE'_{POZ} > EE'_{GDN}$. Although in general we built on the rankings and scores that DMUs can attain for the entire set of feasible weight vectors, these two measures can be used alike the existing DEA ranking methods for assigning a single efficiency score or position to each DMU.
4. *For which DMUs the relative efficiency scores and ranks vary much in the set of feasible weights?* For this purpose, one should analyze the difference between the extreme efficiency scores and ranks as well as the distribution of these measures across all feasible weights. High dispersion of scores and ranks indicates that the priorities of DMUs differ significantly. In some decision contexts, this should prompt investigation as to whether the guidelines for standard practice can be used as a tool to reduce variance in management. In our study, the best example of an airport for which such investigation should be conducted is BZG, which apart from being efficient in the best case, attains efficiency scores lower than 0.5 and ranks in the bottom half for about 25% of feasible weight vectors.
5. *How DMUs perform in one-on-one comparisons?* Traditionally, DEA referred to the efficiency scores and/or ranks. Although these two perspectives are deepened in the proposed framework, the necessary/possible efficiency preference relations and pairwise efficiency outranking indices offer a yet different one-on-one perspective, which is not influenced by the remaining DMUs. Indeed, the analyst may be sometimes more interested in the peer comparison. This is particularly useful if (s)he knows some units better. Then, they can be used as fixed benchmarks

for the remaining DMUs. In our study, an expert interested, e.g., in the performance of SZZ and knowing POZ quite well, would get to know that POZ – despite being inefficient overall – is more efficient than SZZ for all possible priorities assigned to the inputs and outputs.

The necessary preference relation may be very useful also in terms of formulating the corrective actions for the inefficient units. For such units, the efficient ones being necessarily preferred to them represent their hypothetical comparison units (HCUs). Differences in inputs and outputs between DMU and thus identified HCUs clearly indicate the productivity gaps and improvement potential. Moreover, when analyzing the graph of necessary efficiency preference relation, one can think of applying the step-wise benchmarking based on the specification of short-, medium-, and long-term targets. This requires identification of the paths that originate in the node representing some inefficient DMU and finish in one of the nodes corresponding to the efficient unit having no predecessors. In case there are multiple such paths, one may compare the underlying strategies to be potentially adopted. In our study, since $GDN \succeq_E^N KAT \succeq_E^N RZE$ and $BZG \succeq_E^N RZE$, the exemplary recommendation for RZE may be either to follow the example of BZG, or to focus first on reaching the efficiency level of KAT and only then following the practice of GDN.

The robust results can be also applied in other contexts which are important from the managerial perspective:

1. *Specification of performance targets [31,53]:* In the traditional DEA methods, one investigated only the improvement that needs to be made to become efficient (i.e., to be ranked first or to attain the greatest efficiency score for some feasible weight vector). On the contrary, when referring to the robust results, the management may formulate more detailed and diverse questions. In the context of our study, they may concern, e.g., the improvement of performances that warrants that WRO is ranked at worst third for all feasible weights (while currently $R_{WRO,*} = 5 > 3$), or that BZG is necessarily preferred to KAT (while currently $\text{not}(BZG \succeq_E^N KAT)$ and $PEOI'(BZG, KAT) = 99.44\%$), or that the efficiency of WAW is worse at most twice than that of the most efficient unit (while currently $E_{WAW,*} = 0.452 < 0.5$). The answers to these questions can be obtained with LP [53], directly indicating to the management how the DMU's performances should be bettered to attain the desired target.
2. *Identification of outlier DMUs:* The high values for the first rank efficiency acceptability indices and/or efficiency acceptability interval indices for the best scores can be used for detection of the outlier DMUs, similarly as super-efficiencies greater than a pre-defined threshold in a backward approach discussed in [6]. An obvious example of such an outlier in our study is WAW for which $EAI'(WAW, (0.9, 1.0]) = 78.93\%$ and $ERAI'(WAW, 1) = 70.70\%$.
3. *Adding discrimination among the DMUs* by introducing the restrictions on the relative values among different outputs and inputs which represent relevant managerial constraints. This is enhanced by the desirable evolution of the robust results with an incremental specification of weight constraints as discussed in Appendix C.

6. Conclusions

We have proposed an integrated framework for robustness analysis using a data envelopment model. While referring to a ratio-based measure, we considered three different viewpoints on

the efficiency of Decision Making Units in the set of feasible input/output weights. Precisely, we evaluated the units' performance in terms of attained efficiency scores, pairwise preference relations, and ranks. On one hand, we assessed the extreme (in case of scores and ranks), necessary and possible (in case preference relations) performance of units using Linear Programming techniques. On the other hand, we used Monte Carlo simulation to enrich these exact outcomes with stochastic indices. The latter provide estimates of probability of attaining some result as well as some aggregated measures (e.g., expected efficiency score or efficiency rank) derived from a large set of feasible input/output weights. Apart from the complementary characteristics of the considered results, the discussed algorithms compare positively to the traditional techniques of efficiency analysis in terms of requiring less arbitrary assumptions, being less sensitive to a set of considered units, and offering greater discriminative power.

All these benefits have been illustrated on the problem of assessing efficiency of Polish airports. We took into account four inputs (i.e., capacities of a terminal, runways, and an apron, and a catchment area) and two outputs (i.e., passenger traffic and number of aircraft movements) related to the terminal services and movement model. Nevertheless, the scope of problems in which answering similar questions may be of interest to the analyst is very broad. Indeed, our approach can be used in a variety of efficiency analysis problems concerning, e.g., agricultural farms [4], banks [5,34,37,68], container ports and terminals [19,69], courts [54], local governments [21], shipping companies [45], urban rail firms [30], or transportation networks [73].

To support the applicability of our results in other decision contexts, we implemented an open-source software distributed as a part of the *diviz* platform. Apart from providing the modules for both robustness and stochastic analysis, we accounted for the well-known procedures of data envelopment analysis such as super-efficiency or cross-efficiency.

We envisage the following future developments:

- accounting for the hierarchical structure of inputs and outputs [60];
- admitting imprecise performance values;
- extension of the range of considered efficiency preference relations derived from robustness analysis, and studying their properties in terms of transitivity, completeness, reflexivity, continuity, and non-triviality;
- adapting the proposed framework to other data envelopment models such as additive DEA [28,29] or preference models admitting interactions between different inputs and outputs;
- application to different decision problems in transport, medicine, environmental management, and education.

Acknowledgments

The authors thank Sébastien Bigaret and Patrick Meyer from Telecom Bretagne for helping us to make the software available on the *diviz* platform. The work of Miłosz Kadziński and Anna Labijak was supported by the Polish Ministry of Science and Higher Education under the *Iuventus Plus* program (Grant no. IP2015 029674 – 0296/IP5/2016/74).

Appendix A. Alternative formulation of the MILPs for computation of extreme efficiency ranks

To enhance understanding of the underlying reasoning, in this section we present alternative formulations of the MILPs for

computation of the extreme efficiency ranks presented in Section 2.4.1. Instead of minimizing the number of DMUs that can be simultaneously better than DMU_o , R_o^* can be obtained while subtracting from K the cardinality of the maximal subset of DMUs that are at the same time at most as good as DMU_o . For this purpose, the objective function in (7) should be replaced with:

$$\min R_o^* = K - \sum_{k=1, k \neq o}^K b_k,$$

while constraint [*] needs to be substituted with:

$$\sum_{n=1}^N u_n y_{nk} \leq \sum_{m=1}^M v_m x_{mk} + C(1 - b_k) \quad (k = 1, \dots, K, k \neq o).$$

Furthermore, instead of maximizing the number of DMUs that are simultaneously not worse than DMU_o , $R_{*,o}$ can be obtained while subtracting from K the cardinality of the minimal subset of DMUs that are at the same time worse than DMU_o . Then, the objective function in (8) should be replaced with:

$$\max R_{*,o} = K - \sum_{k=1, k \neq o}^K b_k,$$

while constraint [*] needs to be substituted with:

$$\sum_{m=1}^M v_m x_{mk} \leq \sum_{n=1}^N u_n y_{nk} + C b_k \quad (k = 1, \dots, K, k \neq o).$$

Appendix B. Interdependencies between robust results and stochastic indices

The extreme, necessary, and possible results determined with LP influence the stochastic indices in the following way:

Remark B.1. For $DMU_o, DMU_k \in \mathcal{D}$:

1. $i : \{b_{i,*} > E_o^* \vee b_i^* < E_{o,*}\} \Rightarrow EAI'(DMU_o, b_i) = 0$ (i.e., for the efficiency subintervals outside the range delimited by the extreme efficiencies, the efficiency acceptability interval indices are 0, because a unit does not attain such efficiency scores for any feasible weight vector, including these sampled in the Monte Carlo simulation).
2. $\sum_{i: b_{i,*} \leq E_o^* \wedge b_i^* \geq E_{o,*}} EAI'(DMU_o, b_i) = 1$ (i.e., the sum of EAI' 's corresponding to the efficiency intervals with non-empty intersection with $[E_{o,*}, E_o^*]$ is equal to one (see Proposition 2.1 and point 1)).
3. $DMU_o \succ_E^N DMU_k \Rightarrow PEOI'(DMU_o, DMU_k) = 1$ (if the necessary efficiency relation was valid, this needs to be confirmed by all feasible weight vectors, including these sampled in the simulation, and, thus, $PEOI'(DMU_o, DMU_k)$ is equal to one).
4. $\neg(DMU_o \succ_E^P DMU_k) \Rightarrow PEOI'(DMU_o, DMU_k) = 0$ (i.e., if the possible efficiency relation was false, the truth of efficiency preference relation is not confirmed by any feasible weight vector, and, thus, $PEOI'(DMU_o, DMU_k) = 0$).
5. $l : \{l < R_o^* \vee l > R_{o,*}\} \Rightarrow ERAI'(DMU_o, l) = 0$ (i.e., for the ranks outside the interval delimited by the extreme ones, the efficiency rank acceptability indices are 0).
6. $\sum_{l=R_o^*}^{R_{o,*}} ERAI'(DMU_o, l) = 1$ (i.e., the sum of $ERAI'$'s for the ranks between the extreme ones, is equal to one (see Proposition 2.5 and point 5)).

Note that the inverse implications or relations are not necessarily true. In particular, the ranges of efficiencies or ranks determined exactly with LP may be wider than the respective ranges observed in the Monte Carlo sample of weight vectors. Consequently, the estimates EAI' and $ERAI'$ may be equal to 0, whereas the true EAI and $ERAI$ are greater than 0. Further, $PEOI'(DMU_o, DMU_k) = 1$ ($PEOI'(DMU_o, DMU_k) = 0$) does not imply that $DMU_o \succ_E^N$

$DMU_k (\neg(DMU_o \succ_E^p DMU_k))$ since the set of sampled weight vectors might not contain some feasible weights $(v, u) \in (S_v, S_u)$ such that $E_k(v, u) > E_o(v, u)$ ($E_o(v, u) \geq E_k(v, u)$). The only valid interdependencies between the estimates of stochastic indices and extreme, necessary, and possible outcomes are the following:

Remark B.2. For $DMU_o, DMU_k \in \mathcal{D}$:

1. $EAII'(DMU_o, b_i) > 0 \Rightarrow b_{i*} \leq E_o^* \wedge b_i^* \geq E_{o,*}$ (i.e., when an efficiency acceptability interval index is positive, there is at least one feasible weight vectors for which a unit attains efficiency in the respective interval; this implies that the true best efficiency E_o^* of a unit is greater than the lower bound b_{i*} of the interval and its worst efficiency $E_{o,*}$ is less than the upper bound b_i^*).
2. $PEOI'(DMU_o, DMU_k) > 0 \Rightarrow DMU_o \succ_E^p DMU_k$ (i.e. when the pairwise efficiency outranking index is greater than 0, the efficiency preference has been observed for at least one weight vector in the sample, and, thus, the possible efficiency preference relation holds).
3. $PEOI'(DMU_o, DMU_k) = 0 \Rightarrow \neg(DMU_o \succ_E^N DMU_k)$ (i.e. when the pairwise efficiency outranking index is 0, the efficiency preference has not been observed for any weight vector in the sample, and, thus, it certainly does not hold for all feasible weight vectors; this, in turn, implies that the necessary efficiency preference relation is not valid).
4. $ERAI'(DMU_o, l) > 0 \Rightarrow R_o^* \leq l$ and $l \leq R_{o,*}$ (i.e., when an efficiency rank acceptability index for rank l is positive, there is at least one feasible weight vector for which a unit attains l -th position; this implies that the true best efficiency rank R_o^* of the unit and its worst rank $R_{o,*}$ are, respectively, not greater and not less than l).

Appendix C. Evolution of robust results with incremental specification of weight constraints

In this section, we consider a specification of weight constraints in the following iterations of DM's interaction with the proposed framework. We denote with $A_v^1 \subseteq A_v^2 \subseteq \dots \subseteq A_v^s$ nested sets of weight constraints provided by the DM. These sets $A_v^t, t = 1, \dots, s$, generate the respective sets of feasible weight vectors $(S_v, S_u)^t$. These are incrementally constrained, i.e., $(S_v, S_u)^1 \supseteq (S_v, S_u)^2 \supseteq \dots \supseteq (S_v, S_u)^s$. For each iteration $t = 1, \dots, s$, the following results can be derived:

- extreme efficiencies $E_o^{*,t}$ and $E_{o,*}^t$
- possible $\succ_E^{p,t}$ and necessary $\succ_E^{N,t}$ efficiency preference relations,
- extreme efficiency ranks $R_o^{*,t}$ and $R_{o,*}^t$.

The evolution of the robust results with the increase of weight constraints is summarized in Proposition C.1.

Proposition C.1. For $DMU_o \in \mathcal{D}$ and $t = 1, \dots, s-1$:

- $E_o^{*,t} \geq E_o^{*,t+1}$ and $E_{o,*}^t \leq E_{o,*}^{t+1}$ (i.e., in the following iteration, when the space of feasible weights is more constrained, the ranges of attained efficiencies may be narrowed down).
- $\succ_E^{N,t} \subseteq \succ_E^{N,t+1}$ and $\succ_E^{p,t} \supseteq \succ_E^{p,t+1}$ (i.e., the necessary and possible relations may be, respectively, enriched and impoverished).
- $R_o^{*,t} \leq R_o^{*,t+1}$ and $R_{o,*}^t \geq R_{o,*}^{t+1}$ (i.e., the ranking intervals may become narrower, but not wider).

Appendix D. Impact of removing/introducing outlier DMUs on robust results

Traditionally, DEA methods have been focused on identifying the efficient frontier on which the DMUs are considered efficient. In this regard, much attention has been paid to identification of atypical DMUs that may greatly influence the frontier's shape [11,74]. In general, there exist two basic approaches for detection of such outlier

DMUs. On one hand, in a backward approach [6], DMUs with super-efficiencies greater than a pre-defined threshold are identified as outliers. On the other hand, in the forward search procedure [11], the subjectivity of using some arbitrary threshold can be avoided by using a dedicated distance function (see also [10,12]).

In this subsection, we discuss the impact of removing/introducing some (outlier) DMUs on the robust results. In this perspective it is important to remind that all our results are derived from comparing DMUs' efficiencies pairwise rather than measuring their distance from an efficient frontier as in the traditional DEA models.

Let us consider the following subsets of DMUs: $\mathcal{D}' \subset \mathcal{D}'' \subseteq \mathcal{D}$. Thus, $\mathcal{D}'' = \mathcal{D}' \setminus \mathcal{D}'$ contains the DMUs removed from \mathcal{D}' /introduced to \mathcal{D}' . We will denote the results obtained when analyzing a given subset of DMUs by using its symbol in the superscript (e.g., $\succ_E^{N,\mathcal{D}'}$ indicates the necessary efficiency preference relation obtained for \mathcal{D}'). Then, the following proposition summarizes the interdependencies between the outcomes that can be obtained for \mathcal{D}' and \mathcal{D}'' .

Proposition D.1. For $DMU_o, DMU_k \in \mathcal{D}' \subset \mathcal{D}'' \subseteq \mathcal{D}$:

- $EE_o^{\mathcal{D}'} \geq EE_o^{\mathcal{D}''}$ (i.e., for each feasible weight an efficiency attained by DMU_o in \mathcal{D}' is not worse than its respective efficiency in \mathcal{D}'' ; thus, after removing some DMUs, the expected efficiency EE of DMU_o cannot be deteriorated).
- $E_o^{*,\mathcal{D}'} \geq E_o^{*,\mathcal{D}''}$ and $E_{o,*}^{\mathcal{D}'} \geq E_{o,*}^{\mathcal{D}''}$ (i.e., the extreme efficiency scores of each DMU_o obtained within the constrained set \mathcal{D}' are not less than its respective scores within \mathcal{D}'').
- $PEOI^{\mathcal{D}'}(DMU_o, DMU_k) = PEOI^{\mathcal{D}''}(DMU_o, DMU_k)$ (although the absolute efficiency scores attained by DMU_o and DMU_k may change after removing/introducing some other DMUs, the order between these scores remains the same for all feasible weight vectors; consequently, the value of pairwise efficiency outranking index for a given pair of DMUs does not depend on the remaining units, being the same in both \mathcal{D}' and \mathcal{D}'').
- $DMU_o \succ_E^{N,\mathcal{D}'} DMU_k \Leftrightarrow DMU_o \succ_E^{N,\mathcal{D}''} DMU_k$ (the above justification proves that the status of \succ_E^N for a given pair of DMUs does not depend on other units; as a result, for all pairs of units contained in $\mathcal{D}' \times \mathcal{D}'$, the truth or falsity of \succ_E^N is the same in both \mathcal{D}' and \mathcal{D}'').
- $0 \leq ER_o^{\mathcal{D}'} - ER_o^{\mathcal{D}''} \leq |\mathcal{D}''|$ (i.e., for each feasible weight vector DMU_o is ranked not worse in \mathcal{D}' than in \mathcal{D}'' ; in fact, it can be ranked better by at most $|\mathcal{D}''|$, which is the cardinality of the removed subset of DMUs; thus, after removing some DMUs from \mathcal{D}' , the expected rank ER of DMU_o cannot be deteriorated, being at most by $|\mathcal{D}''|$ better (lower) in \mathcal{D}' than in \mathcal{D}'').
- $0 \leq R_o^{*,\mathcal{D}'} - R_o^{*,\mathcal{D}''} \leq |\mathcal{D}''|$ and $0 \leq R_{o,*}^{\mathcal{D}'} - R_{o,*}^{\mathcal{D}''} \leq |\mathcal{D}''|$ (i.e., after removing $|\mathcal{D}''|$ units from \mathcal{D}' , the extreme ranks of DMU_o in \mathcal{D}' cannot be deteriorated, being at most by $|\mathcal{D}''|$ better in \mathcal{D}' than in \mathcal{D}'').

Since the efficiency acceptability interval indices $EAIIs$ and efficiency rank acceptability indices $ERAIIs$ depend on the entire set of DMUs and all feasible sets of weights, one cannot formulate any general remarks for their evolution after removing/introducing some DMUs.

References

- [1] Adler N, Berechman J. Measuring airport quality from the airlines' viewpoint: an application of data envelopment analysis. *Transport Policy* 2001;8:171–81.
- [2] Andersen P, Petersen N. A procedure for ranking efficient units in DEA. *Management Science* 1993;39:1261–4.
- [3] Athanassopoulos A, Podinovski V. Dominance and potential optimality multiple criteria models decision analysis with imprecise information. *Journal of the Operational Research Society* 1997;48:142–50.
- [4] Atici K, Podinovski V. Using data envelopment analysis for the assessment of technical efficiency of units with different specialisations: an application to agriculture. *Omega* 2015;54:72–83.
- [5] Avkiran N. An illustration of dynamic network DEA in commercial banking including robustness tests. *Omega* 2015;55:141–50.
- [6] Banker R, Chang H. The super-efficiency procedure for outlier identification, not for ranking efficient units. *European Journal of Operational Research* 2006;175(2):1311–20.

- [7] Barros C. Technical change and productivity growth in airports: a case study. *Transportation Research, Part A* 2008;42:818–32.
- [8] Barros C, Dieke P. Measuring the economic efficiency of airports: a Simar–Wilson methodology analysis. *Transportation Research Part E* 2008;44:1039–51.
- [9] Barros C, Weber W. Productivity growth and biased technological change in UK airports. *Transportation Research Part E* 2009;45:642–53.
- [10] Bellini T. Detecting atypical observations in financial data: the forward search for elliptical copulas. *Advances in Data Analysis and Classification* 2010;4(4):287–99.
- [11] Bellini T. Forward search outlier detection in data envelopment analysis. *European Journal of Operational Research* 2012;216(1):200–7.
- [12] Bellini T. The forward search interactive outlier detection in cointegrated var analysis. *Advances in Data Analysis and Classification* 2015:1–23.
- [13] Bouyssou D. Using DEA as a tool for MCDM. *Journal of the Operational Research Society* 1999;50:974–8.
- [14] Charnes A, Cooper W, Rhodes E. Measuring the efficiency of decision making units. *European Journal of Operational Research* 1978;2(6):429–44.
- [15] Chinneck J. Feasibility and infeasibility in optimization: algorithms and computational methods. New York: Springer; 2008.
- [16] Civil Aviation Authority, 2009. Analysis of the Polish airline market (in Polish), Warsaw, Poland.
- [17] Cook W, Tone K, Zhu K. Data envelopment analysis: prior to choosing a model. *Omega* 2014;44:1–4.
- [18] Cooper W, Seiford L, Zhu J. Handbook on data envelopment analysis. International series in operations research & management science. New York: Springer; 2011.
- [19] Cullinane K, Wang T-F, Song D-W, Ji P. The technical efficiency of container ports: comparing data envelopment analysis and stochastic frontier analysis. *Transportation Research Part A: Policy and Practice* 2006;40(4):354–74.
- [20] Dias L, Climaco J. Aiding decisions with multiple criteria—essays in honor of Bernard Roy. In: Bouyssou D, Jacquet-Lagréze E, Perny P, Slowiński R, Vanderpooten D, Vincke P, editors. New advances in multiple criteria decision analysis. Dordrecht: Kluwer; 2002. p. 175–93.
- [21] Doumpos M, Cohen S. Applying data envelopment analysis on accounting data to assess and optimize the efficiency of greek local governments. *Omega* 2014;46:74–85.
- [22] Doyle J, Green R. Efficiency and cross-efficiency in DEA: derivations, meanings and uses. *Journal of the Operational Research Society* 1994;45:567–78.
- [23] Emrouznejad A, Barnett B, Tavares G. Evaluation of research in efficiency and productivity: a survey and analysis of the first 30 years of scholarly literature in DEA. *Socio-Economic Planning Sciences* 2008;42:151–7.
- [24] Fernandes E, Pacheco RR. Efficient use of airport capacity. *Transportation Research : Part A. Policy and Practice* 2002;36:225–38.
- [25] Fung M, Wan K, Hui Y, Law J. Productivity changes in Chinese airports 1995–2004. *Transportation Research Part E* 2008;44:521–42.
- [26] Gillen D, Lall A. Developing measures of airport productivity and performance: an application of data envelopment analysis. *Transportation Research Part E: Logistics and Transportation Review* 1997;33:261–73.
- [27] Gillen D, Lall A. Non-parametric measures of efficiency of US airports. *International Journal of Transport Economics* 2001;28:283–306.
- [28] Gouveia M, Dias L, Antunes C. Additive DEA based on MCDA with imprecise information. *Journal of the Operational Research Society* 2008;59:54–63.
- [29] Gouveia M, Dias L, Antunes C, Boucinha J, Inácio C. Benchmarking of maintenance and outage repair in an electricity distribution company using the value-based DEA method. *Omega* 2015;53:104–14.
- [30] Graham D. Productivity and efficiency in urban railways: parametric and non-parametric estimates. *Transportation Research Part E: Logistics and Transportation Review* 2008;44(1):84–99.
- [31] Kadziński M, Ciomek K, Rychy P, Slowiński R. Post factum analysis for robust multiple criteria ranking and sorting. *Journal of Global Optimization*; 2016. p. 1–32 (in press). <http://dx.doi.org/10.1007/s10898-015-0359-3>.
- [32] Kadziński M, Tervonen T. Robust multi-criteria ranking with additive value models and holistic pair-wise preference statements. *European Journal of Operational Research* 2013;228(1):169–80.
- [33] Kadziński M, Tervonen T. Stochastic ordinal regression for multiple criteria sorting problems. *Decision Support Systems* 2013;55(1):55–66.
- [34] Kao C, Liu S-T. Multi-period efficiency measurement in data envelopment analysis: the case of Taiwanese commercial banks. *Omega* 2014;47:90–8.
- [35] Kocak H. Efficiency examination of Turkish airports with DEA approach. *International Business Research* 2011;4:204–12.
- [36] Lahdelma R, Salminen P. Stochastic multicriteria acceptability analysis using the data envelopment model. *European Journal of Operational Research* 2006;170(1):241–52.
- [37] LaPlante A, Paradi J. Evaluation of bank branch growth potential using data envelopment analysis. *Omega* 2015;52:33–41.
- [38] Liu J, Lu L, Lu W-M. Research fronts in data envelopment analysis. *Omega* 2016;58:33–45.
- [39] Liu J, Lu L, Lu W-M, Lin B. Data envelopment analysis 1978–2010: a citation-based literature survey. *Omega* 2013;41(1):3–15.
- [40] Liu J, Lu L, Lu W-M, Lin B. A survey of DEA applications. *Omega* 2013;41(5):893–902.
- [41] Martini J, Roman C. An application of DEA to measure the efficiency of Spanish airports prior to privatization. *Journal of Air Transport Management* 2001;7(3):149–57.
- [42] Meyer P, Bigaret S. Diviz: a software for modeling, processing and sharing algorithmic workflows in MCDA. *Intelligent Decision Technologies* 2012;6:283–96.
- [43] Mousseau V, Figueira J, Dias L, da Silve CG, Climaco J. Resolving inconsistencies among constraints on the parameters of an MCDA model. *European Journal of Operational Research* 2003;147(1):72–93.
- [44] Murillo-Melchor C. An analysis of technical efficiency and productive change in Spanish airports using the Malmquist index. *International Journal of Transport Economics* 1999;26:271–92.
- [45] Panayides P, Lambertides N, Savva C. The relative efficiency of shipping companies. *Transportation Research Part E: Logistics and Transportation Review* 2011;47(5):681–94.
- [46] Parker D. The performance of the BAA before and after privatisation. *Journal of Transport Economics and Policy* 1999;33:133–46.
- [47] Pels E, Nijkamp P, Rietveld P. Relative efficiency of European airports. *Transport Policy* 2001;8:183–92.
- [48] Pels E, Nijkamp P, Rietveld P. Inefficiencies and scale economies of European airport operations. *Transportation Research Part E* 2003;39:341–61.
- [49] Podinovski V. DEA models for the explicit maximisation of relative efficiency. *European Journal of Operational Research* 2001;131(3):572–86.
- [50] Podinovski V. The explicit role of weight bounds in models of data envelopment analysis. *Journal of the Operational Research Society* 2005;56:1408–18.
- [51] PriceWaterhouseCoopers, Oliver Wyman, MKMetric GmbH, Deutsche Flugsicherung GmbH. Concept of the central airport in Poland (in Polish), Polish Ministry of Infrastructure, Warsaw, Poland; 2010.
- [52] Roy B. Robustness in operational research and decision aiding: a multi-faceted issue. *European Journal of Operational Research* 2010;200(3):629–38.
- [53] Salo A, Punkka A. Ranking intervals and dominance relations for ratio-based efficiency analysis. *Management Science* 2011;57:200–14.
- [54] Santos S, Amado C. On the need for reform of the Portuguese judicial system—Does data envelopment analysis assessment support it? *Omega* 2014;47:1–16.
- [55] Sarkis J. A comparative analysis of DEA as a discrete alternative multiple criteria decision tool. *European Journal of Operational Research* 2000;123:543–57.
- [56] Sarkis J. Operational efficiency of major US airports. *Journal of Operations Management* 2000;18:335–51.
- [57] Sarkis J, Talluri S. Performance-based clustering for benchmarking of US airports. *Transportation Research Part A* 2004;38:246–329.
- [58] Seiford L, Zhu J. Stability regions for maintaining efficiency in data envelopment analysis. *European Journal of Operational Research* 1998;108:127–39.
- [59] Sexton T, Silkman R, Hogan A. Data envelopment analysis: critique and extensions. In: Silkman RH, editor. Measuring the efficiency: an assessment of data envelopment analysis. American evaluation association. San Francisco: Jossey Bass Inc.; 1986. p. 73–105.
- [60] Shen Y, Hermans E, Brijts T, Wets G. Data envelopment analysis for composite indicators: a multiple layer model. *Social Indicators Research* 2013;114(2):739–56.
- [61] Stewart T. Relationships between DEA and MCDM. *Journal of the Operational Research Society* 1996;47:654–65.
- [62] Tervonen T, van Valkenhoef G, Basturk N, Postmus D. Hit-and-run enables efficient weight generation for simulation-based multiple criteria decision analysis. *European Journal of Operational Research* 2013;224(3):552–9.
- [63] Thompson R, Langemeier LN, Lee C, Lee E, Thrall R. The role of multiplier bounds in efficiency analysis with application to Kansas farming. *Journal of Econometrics* 1990;46:93–108.
- [64] Tsou C-M, Huang D-Y. On some methods for performance ranking and correspondence analysis in the DEA context. *European Journal of Operational Research* 2010;203(3):771–83.
- [65] Tsui W, Balli H, Gilbey A, Gow H. Operational efficiency of Asia-Pacific airports. *Journal of Air Transport Management* 2014;40:16–24.
- [66] van Valkenhoef G, Tervonen T, Postmus D. Notes on 'hit-and-run enables efficient weight generation for simulation-based multiple criteria decision analysis'. *European Journal of Operational Research* 2014;239(3):865–7.
- [67] Vincke P. Robust solutions and methods in decision-aid. *Journal of Multi-Criteria Decision Analysis* 1999;8(3):181–7.
- [68] Wang K, Huang W, Wu J, Liu Y-N. Efficiency measures of the Chinese commercial banking system using an additive two-stage DEA. *Omega* 2014;44:5–20.
- [69] Wu Y-CJ, Goh M. Container port efficiency in emerging and more advanced markets. *Transportation Research Part E: Logistics and Transportation Review* 2010;46(6):1030–42.
- [70] Yang H-H. Measuring the efficiencies of Asia-Pacific international airports - parametric and non-parametric evidence. *Computers & Industrial Engineering* 2010;59(4):697–702.
- [71] Yoshida Y, Fujimoto H. Japanese-airport benchmarking with DEA and endogenous-weight TFP methods: testing the criticism of over-investment in Japanese regional airports. *Transportation Research Part E* 2004;40:533–46.
- [72] Yu M-M, Chern C-C, Hsiao B. Human resource rightsizing using centralized data envelopment analysis: evidence from Taiwan's airports. *Omega* 2013;41(1):119–30.
- [73] Zhao Y, Triantis K, Murray-Tuite P, Edara P. Performance measurement of a transportation network with a downtown space reservation system: a network-DEA approach. *Transportation Research Part E: Logistics and Transportation Review* 2011;47(6):1140–59.
- [74] Zhu J. Robustness of the efficient DMUs in data envelopment analysis. *European Journal of Operational Research* 1996;90:451–60.
- [75] Zhu J. Super-efficiency and DEA sensitivity analysis. *European Journal of Operational Research* 2001;129:443–55.

Publication [P2]

P. Gasser, M. Cinelli, A. Labijak, M. Spada, P. Burgherr, M. Kadziński, and B. Stojadinović. Quantifying electricity supply resilience of countries with robust efficiency analysis. *Energies*, 13(7):1535, 2020, DOI: 10.3390/en13071535.

Number of citations²:

- according to Web of Science: 7
- according to Google Scholar: 9

Contribution of the author of this dissertation and four co-authors:

- Anna Labijak-Kowalska
 - Co-authorship of the idea of application of robustness analysis methods for DEA in the study of analysis of the resilience of countries' electricity systems,
 - application of the ratio-based DEA to obtain the sets of efficient and inefficient countries, projections onto the efficient frontier, and necessary improvements for inefficient countries,
 - application of the robustness analysis methods to the data considered in the study,
 - implementation of the software necessary in the study,
 - authorship of the concept and algorithms for determination of efficiency reducts and costructs,
 - preparation of the updated results for the three scenarios for countries development,
 - consultations on the data preparation and interpretation of the results,
 - reviewing and correcting the text of the manuscript in terms of the methodology of the analysis.
- Patrick Gasser
 - Collection of the data and in the considered case study,
 - selection of the indicators for the case study,
 - co-authorship of the text of the publication in the application-oriented parts,
 - interpretation and discussion of the results.
- Marco Cinelli
 - Co-authorship of the idea underlying the paper consisting of applying the robust efficiency method to the considered study,
 - collection and preparation of the data for the considered study,

²as on May 30, 2023

- interpretation and discussion of the results,
- reviewing and corrections on the text of the publication.
- Matteo Spada
 - Co-authorship of the idea underlying the paper consisting of analyzing electricity supply resilience of countries,
 - supervision of the data collection and indicator selection process for the considered study,
 - reviewing and correcting the text of the manuscript.
- Peter Burgherr
 - Overall coordination as responsible Principal Investigator within the Future Resilient Systems (FRS) program of the Singapore ETH Centre (SEC),
 - co-authorship of the idea underlying the paper consisting of analyzing electricity supply resilience of countries
 - supervision of the data collection and indicator selection process for the considered study;
 - reviewing and correcting the text of the manuscript.

Article

Quantifying Electricity Supply Resilience of Countries with Robust Efficiency Analysis

Patrick Gasser ^{1,*}, Marco Cinelli ^{1,†}, Anna Labijak ², Matteo Spada ³, Peter Burgherr ³, Miłosz Kadziński ² and Božidar Stojadinović ⁴

¹ Future Resilient Systems (FRS), Singapore-ETH Centre (SEC), Swiss Federal Institute of Technology (ETH) Zürich, Singapore 138602, Singapore; marco.cinelli@put.poznan.pl

² Institute of Computing Science, Poznań University of Technology (PUT), 60-965 Poznań, Poland; anna.labijak@cs.put.poznan.pl (A.L.); milosz.kadzinski@cs.put.poznan.pl (M.K.)

³ Laboratory for Energy Systems Analysis (LEA), Paul Scherrer Institut (PSI), 5232 Villigen PSI, Switzerland; matteo.spada@psi.ch (M.S.); peter.burgherr@psi.ch (P.B.)

⁴ Department of Civil, Environmental and Geomatic Engineering, Institute of Structural Engineering, Swiss Federal Institute of Technology (ETH) Zürich, 8093 Zurich, Switzerland; stojadinovic@ibk.baug.ethz.ch

* Correspondence: patrick.gasser@frs.ethz.ch

† Present address: Institute of Computing Science, Poznań University of Technology, 60-965 Poznań, Poland.

Received: 30 January 2020; Accepted: 16 March 2020; Published: 25 March 2020



Abstract: The interest in studying energy systems' resilience is increasing due to a rising awareness of the importance of having a secure energy supply. This growing trend is a result of a series of recent disruptions, among others also affecting electricity systems. Therefore, it is of crucial importance for policymakers to determine whether their country has a resilient electricity supply. Starting from a set of 12 indicators, this paper uses data envelopment analysis (DEA) to comprehensively evaluate the electricity supply resilience of 140 countries worldwide. Two DEA models are applied: (1) the original ratio-based Charnes, Cooper, and Rhodes (CCR) model and (2) a novel hybrid framework for robust efficiency analysis incorporating linear programming and Monte Carlo simulations. Results show that the CCR model deems 31 countries as efficient and hence lacks the capability to differentiate them. Furthermore, the CCR model considers only the best weight vectors for each country, which are not necessarily representative of the overall performance of the countries. The robustness analysis explores these limitations and identifies South Korea, Singapore and Canada as the most resilient countries. Finally, country analyses are conducted, where Singapore's and Japan's performances and improvement potentials are discussed.

Keywords: data envelopment analysis; electricity supply; resilience; energy security; ratio-based efficiency model; robustness analysis

1. Introduction

Electricity is a crucial commodity to foster the economic development and well-being of a country [1]. Governments are increasingly aware of the need to improve the energy efficiency of electricity production (i.e., decrease the total amount of energy required to produce the desired quantity of electricity), as it leads to better supply security and reduced greenhouse gas emissions. Even though recently the global average efficiency slowly improved [1], major electricity supply disruptions still happen (e.g., the 2012 India blackout [2] or the 2015 Turkey blackout [3]). The financial, social and environmental consequences of such disruptions can cause great damage to the economy of a country, its government and citizens [4,5]. Resilience aims at minimizing the impact of these adverse consequences by defining pre- and post-event strategies, making outages less likely or smaller in extent [6,7].

A resilient electricity supply is fundamental to guarantee a well-functioning modern society [8]. In this regard, one of the key interests of policymakers is to assess how their country performs compared to others. This kind of international comparison supports them in identifying improvement potentials and quantifying achievements or progress towards predefined objectives and targets [9]. Many international country performance assessments are indicator-based [10], because indicators are suitable to model multi-dimensional problems [11]. Countries are commonly ranked based on the indicators only or on aggregated measures, sometimes also called indices or composite indicators, that combine the individual indicators [12]. Within the energy sector, a wide range of operational research methods are used to build such indices [13]. Examples include the analytical hierarchy process (AHP) [14,15], technique for order preference by similarity to ideal solution (TOPSIS) [16], outranking methods such as the preference ranking organization method for enrichment evaluation (PROMETHEE) [17] or elimination et choix traduisant la réalité (ELECTRE) [18], weighted averages [19], multi-attribute value/utility theory (MAVT/MAUT) [20] and data envelopment analysis (DEA) [21].

In the context of electricity supply resilience, comparative country evaluations and rankings are missing [22]. Hence, building upon the framework proposed by Gasser et al. (2017) [23] and Gasser et al. (2020) [24] to define a set of 12 indicators that cover resilience holistically, this paper assesses the electricity supply resilience of 140 countries following a security of supply perspective. Due to the fact that the indicators have positive and negative preference orders, and the requirement to not involve preferences from decision-makers, DEA, a family of non-parametric methods to derive efficiencies of decision-making units (DMUs) [25], is particularly suitable to rank the countries. In fact, for DEA, the indicator weights are endogenously determined (directly calculated from the performance matrix itself) [26]. The final scores of the DMUs are commonly called efficiencies, as they represent the ratio of the weighted output indicators' performances to the weighted input indicators performances. Hence, a DMU that is deemed efficient is also resilient, as the indicator set represents electricity supply resilience. DEA efficiency and resilience have thus equal meanings in this context and the terms can be used interchangeably.

The DEA methodology hereby developed allows the following important questions to be answered, which are relevant to (inter-) governmental agencies as well as research institutions:

1. What are the best performing (i.e. most resilient) countries and what are the reasons for this achievement (see Section 4.1)?
2. Why are some countries inefficient, how can they improve their scores and which are their benchmarks (see Section 4.2)?
3. How robust is the performance of the countries (see Section 4.3)?
4. What is the univocal ranking of the countries (see Section 4.4)?
5. How well does a country perform in comparison to another one (see Section 4.5)?
6. How does the performance of countries vary according to changes in selected indicators (see Section 4.6)?

The current paper is organized as follows. In Section 2, a detailed literature review about energy-related country comparisons using DEA is provided, which leads to the formulation of the research gaps. Section 3 describes the case study and the methodology. The latter includes the original ratio-based Charnes, Cooper, and Rhodes (CCR) DEA model [27] and a hybrid framework for robust efficiency analysis incorporating linear programming and Monte Carlo simulations [28]. The CCR model was considered because it is the most commonly used DEA model. The hybrid framework explores the limitations of the CCR model by performing a robustness assessment through selecting random weight vectors obtained via a Hit-And-Run algorithm. To the authors' best knowledge, the present study represents the first application of such an analysis to a country ranking. In Section 4, comparative results answering the research questions are presented and discussed. Furthermore, improvement potentials for Singapore and Japan are analyzed. Section 5 provides the main conclusions of the study.

2. Literature Review—Energy-Related Country Comparisons with Data Envelopment Analysis

The first notions of DEA can be traced back to early publications by Farrell (1957) [29] and Brockhoff (1970) [30], but the seminal paper by Charnes et al. (1978) provided the first application of linear programming to estimate an efficiency frontier [27]. A comprehensive overview of DEA in past decades is given in a review by Liu et al. (2013) [31]. The popularity of DEA is also reflected by its numerous applications in different fields such as the evaluation of socio-economic, environmental and productivity performance [31]. There are also diverse applications in the energy sector [32–37], as shown in Table 1.

A detailed, global study by Wang (2015) compared the sustainability of the energy systems in 109 countries [38]. However, this study was based on only three, rather generic, indicators from the World Bank: (1) the CO₂ emissions intensity in kg per 2005 USD of Gross Domestic Product (GDP), (2) the energy intensity in kg of oil equivalent per GDP (constant 2005 Purchasing Power Parity (PPP)), and (3) the share of electricity produced from renewables in %. Overall, results demonstrated that the energy systems in high-income countries have a better sustainability performance.

Another worldwide study analyzes emission reductions, energy conservation and economic output of 87 countries based on the GDP, the capital stock, the labor force, the energy consumption and the overall CO₂ emissions [39]. European countries were found to perform better than non-European ones. Li and Wang (2014) used the same indicators and applied them to 95 countries [40]. They identified tremendous gaps between countries according to income groups, and similarly high-income countries were ranked top.

Table 1. Literature review of publications using data envelopment analysis (DEA) to analyze energy-related topics.

Source	Scope	Geographical Coverage	Number of DMUs	Inputs	Outputs
Apergis et al. (2015) [34]	Energy efficiency	Organisation for Economic Co-operation and Development (OECD)	20	Labor; energy consumption; capital stock	GDP; CO ₂ emissions
Bampatsou et al. (2013) [41]	Capacity of an economy to produce a higher GDP given fixed energy inputs	European Union	15	Fossil and non-fossil fuel energy consumption	GDP
Cai et al. (2019) [42]	Carbon emissions efficiency in Chinese cities	China	280	Labor; capital; energy and water consumption	GDP; CO ₂ emissions
Camarero et al. (2013) [43]	Impact of CO ₂ , SO ₂ and NOx air-pollutants on the environment	OECD	22	CO ₂ , SO ₂ and NOx emissions	GDP
Chang (2014) [44]	Energy intensity	European Union	27	Capital stock; labor force; energy consumption	GDP
Cui et al. (2014) [45]	Energy efficiency	Global	9	Employees; energy consumption; energy services	CO ₂ emissions; industrial profit
Gómez-Calvet et al. (2016) [46]	Abatement opportunities of CO ₂ , SO ₂ and NOx air-pollutants	European Union	27	CO ₂ , SO ₂ and NOx emissions	GDP
Halkos and Petrou (2019) [47]	Energy recovery from waste	European Union	28	Energy consumption; labor; capital; population density	GDP; Greenhouse Gases (GHG), NOx and SOx emissions
Hsieh et al. (2019) [48]	Environmental assessment	European Union	28	Labor; capital; energy consumption	GHG and SOx emissions; GDP
Hu and Kao (2007) [49]	Energy-saving target ratio	Asia-Pacific	17	Energy; labor; capital	GDP
Li and Wang (2014) [40]	Environmental efficiency	Global	95	Capital stock; labor force; energy consumption	GDP; CO ₂ emissions
Liou and Wu (2011) [50]	Effect of economic development on energy use efficiency and CO ₂ emissions	Global	57	Labor; capital; energy consumption	GDP; CO ₂ emissions
Pang et al. (2015) [39]	Clean energy use and total-factor efficiencies	Global	87	Capital stock; labor force; energy consumption	GDP; CO ₂ emissions
Ramanathan (2005) [51]	Energy consumption and carbon dioxide emissions	Middle East and North Africa	17	Fossil fuel energy consumption; carbon emissions	Non-fossil fuel energy consumption; GDP

Table 1. Cont.

Source	Scope	Geographical Coverage	Number of DMUs	Inputs	Outputs
Robaina-Alves et al. (2015) [52]	Resource and environment efficiency	Europe	26	Capital stock; labor force; energy consumption	GDP; GHG emissions
Song et al. (2013) [53]	Energy efficiency	Brazil, Russia, India, China and South Africa (BRICS)	5	Energy consumption; economically active population; capital	GDP
Wang et al. (2019) [54]	Relation between CO ₂ emissions and GDP	Global	25	Gross capital formation; labor force; energy consumption	GDP; CO ₂ emissions
Wang (2015) [38]	Energy systems' sustainability	Global	109	CO ₂ emissions; energy intensity	Share of renewables
Wegener and Amin (2019) [37]	Greenhouse gas emissions minimization	Canada and USA	23	Wells; employees; capital expenditures; total assets	GHG emissions; production
Zeng et al. (2017) [55]	Economic; energy supply; environmental	Baltic States	3	Energy intensity; energy weight in HICP; electricity prices; import dependency; diversification of import sources; diversification of energy mix	Energy balance of trade; share of renewables; carbon intensity
Zhang et al. (2011) [56]	Total-factor energy efficiency	Developing countries	23	Labor force; energy consumption; capital stock	GDP
Zhou and Ang (2008) [57]	Energy efficiency performance	OECD	21	Capital stock; labor force; consumption of coal, oil, gas and other	GDP; CO ₂ emissions
Zhou et al. (2014) [58]	Energy efficiency of transport sector	China	30	Labor; consumption of coal, gasoline, kerosene, diesel oil, electricity and other	Passenger kilometers; tonne-kilometers; CO ₂ emissions
Zhou et al. (2016) [59]	Energy efficiency	Global	32	Capital stock; labor force; fossil and non-fossil energy consumption	GDP; CO ₂ emissions
This study	Electricity supply resilience	Global	140	System Average Interruption Duration Index (SAIDI); accident risks; import dependence; average outage time	Control of corruption; political stability and absence of violence/terrorism; mix diversity; equivalent availability factor; GDP per capita; insurance penetration; government effectiveness; ease of doing business

Gómez-Calvet et al. (2016) studied the opportunities for abatement of CO₂, SO₂ and NO_x air-pollutants by looking at the evolution of environmental performance over time [46]. They found that the environmental efficiency of European countries had improved over the period 1993–2010. The same three air-pollutants were also analyzed by Camarero et al. (2013) and they reached similar conclusions, with the exception that the eco-efficiency of NO_x emissions did not improve [43]. Regarding energy efficiency improvements, Wang et al. (2019) compared the CO₂ emissions in relation to GDP growth from 25 countries and found that India and China are the two worst countries in terms of energy efficiency [54]. A similar study focusing on energy recovery from waste for European Union (EU) member states was produced by Halkos and Petrou (2019) [47]. Furthermore, Robaina-Alves et al. (2015) derived the efficiencies of European countries based on the maximization of the ratio between the GDP (desired output) and GHG emissions (undesired output) [52]. Their key finding was that since the ratification of the Kyoto Protocol, countries have taken steps to reduce emissions and this is reflected in the evolution of the eco-efficiency level of some countries.

Several other publications address energy-efficiency issues in general. Cui et al. (2014) found that energy efficiency is mostly driven by investments into energy technologies research and tax exemptions for technology companies [45]. Based on the analysis of 28 European countries, Hsieh et al. (2019) recommend that “the EU’s strategy for environmental energy improvement should be to pay attention to the benefits of renewable energy utilization, reducing GHG emissions, and enhancing the development of renewable energy utilization to help achieve the goal of lower GHG emissions” [48]. Apergis et al. (2015) show that capital-intensive countries are more energy efficient than labor-intensive ones [34]. Zhang et al. (2011) compare 23 developing countries according to their total-factor energy efficiency, which is defined as the ratio between the targeted energy input and the actual energy input [56]. Similarly, Chang (2014) studies the difference between the targeted and actual energy intensities of 27 EU member countries in order to make conclusions about the potentials for improvement [44]. With a data set of 57 countries, Liou and Wu (2011) found that economic development is interrelated with energy use efficiency and CO₂ emission control [50]. Furthermore, Zhou et al. (2014) used the DEA to rank 30 administrative regions of China according to the energy efficiency of their transport sector [58]. Results show that the Eastern area generally performs better than the Central and Western areas. Song et al. (2013) found that the economies of Brazil, Russia, India, China and South Africa (BRICS) have low energy efficiencies but the trend is increasing quickly [53]. Cai et al. (2019) quantify the carbon emissions of 280 Chinese cities to find that only nine of them are efficient [42]. Results show that coastal regions are performing better than central and western regions. Finally, Zhou et al. (2016) develop novel energy-efficiency measures that seem to handle undesirable outputs better and are more effective at identifying inefficient production behaviour [60]. Their data set consists of 32 countries.

Economic efficiency has also been analyzed as the capacity of an economy to produce higher GDP for a given total energy input. In particular, Bampatsou et al. (2013) study the effect of different energy mixes and find that adding nuclear energy into a country’s energy mix affects negatively its economic efficiency, due to fewer efforts invested in energy saving and conservation [41].

In summary, as seen in Table 1, most of these studies deal with energy or eco-efficiency. Out of the 25 studies analyzed (including the present one), 15 of them (60%) use the same set of inputs: labor force, capital stock and energy consumption [34,39,40,42,44,47–50,52–54,56,57,59]. Furthermore, out of these 15 studies, four consider only the GDP as an output [44,49,53,56], while the rest consider the GDP as a desirable output and GHG emissions as an undesirable output [34,39,40,42,47,48,50,52,54,57,59]. Within DEA country comparisons, efficiency is, therefore, usually measured as a minimization of the labor force, capital stock, and energy consumption (inputs) in order to maximize the GDP (desirable output) and minimize the GHG emissions (undesirable output). Further studies use only a subset of these indicators, such as only the energy consumption as an input and the GDP as an output [41], or only the GHG emissions as an input and the GDP as an output [43,46]. Overall, only five studies do not directly use the GDP as an output [37,38,45,55,58]. Three studies differentiate between fossil fuel and non-fossil fuel energy consumption [38,51,55]. The fossil fuel energy consumption would

be classified as an input (hence to be minimized), while the non-fossil fuel or renewable sources are classified as outputs (hence to be maximized).

Finally, to the authors' best knowledge, no study using DEA to comprehensively rank countries with regard to their electricity supply resilience performance has been published. This implies that a broader set of indicators needs to be considered, including grid reliability, accident risks, diversity of generation and availability of production technologies. Furthermore, the present study uses DEA models to assess ranking robustness, and to explore scenario-based improvement strategies. Therefore, the research gaps filled by the present study are:

1. The use of DEA models to develop rankings that represent the electricity supply resilience of countries.
2. The development of novel DEA algorithms to better understand why some countries are efficient and others are not. These new models are applied for the first time in a real-life case study.
3. The examination of ranking stability by means of robustness analysis.
4. The study of country-specific improvement strategies from an optimization point of view.

3. Case Study Description and Methodology

In this section, first the general scope of the case study and the indicator set selection process are described. Second, the quantification of the data and preparation for DEA are presented. Finally, the DEA concept, its notation and formulas are explained.

3.1. Indicator Set Selection, Quantification and Data Set Preparation

The first step was to conduct a literature review in order to identify relevant indicators. As shown by Gasser et al. (2017) [23], most of the indicators used in the security of electricity supply studies are related to resilience too. Hence, the abundance of security of supply studies is a promising starting point (e.g., [10,61–65]). On top of these, further ones related to resilience were considered as well (e.g., [8,66]). This resulted in an extensive list of resilience-related indicators. Subsequently, all of these were assessed according to four quality criteria [67]: (1) their relevance to resilience, (2) the credibility of the data source, (3) the availability of the data and (4) the comparability of the data between countries. The more countries in the list, the more likely data becomes unavailable or incomparable. Thus, as the present case study is about the electricity sector, the limiting data was the electricity production by fuel type. The most comprehensive data for this comes from the International Energy Agency (IEA) and is available for 140 countries worldwide [68]. Therefore, the final set of 140 countries consists of 12 indicators fulfilling the four assessment criteria (see Table 2) [24,69]. The indicators cover resilience holistically, i.e., both the pre- and post-event phases, and represent, among others, the quality, reliability and interconnectivity of the electricity system, the generation diversity, the fuel supply security and self-sufficiency, the available financial resources, the equivalent availability factor (EAF), the average outage times and geopolitical factors such as corruption, government effectiveness and political stability. Overall, this study covers more than 96% of the world's population and 99.6% of the world's electricity consumption.

The data come from a variety of reliable, credible and widely recognized sources, ranging from governmental agencies to international organizations and private companies [68,70–75]. Furthermore, the values of each indicator originate from a unique source, making it homogeneous and comparable. However, there are some missing values, which were inserted as the mean of the other values in order not to distort the data [76]. Inserting the mean of the other values is one of the three most standard techniques employed for dealing with data incompleteness in different scientific disciplines [77]. The other two procedures consist in (1) excluding the incomplete cases and (2) replacing the unknown value with the entire range of all possible values on a given indicator (input or output). These two other procedures were neglected for the following reasons. First, the countries with missing performances were not excluded from the analysis as the data was not available only for a limited subset of inputs

or outputs, whereas for the vast majority of countries it has been reliably collected. In the context of DEA, the practical usefulness and credibility of the results increase with the greater number of DMUs. Considering that we have 140 countries for only 12 indicators, the ratio country to indicators is very high [32,78]. This increases the discriminatory power of DEA, ultimately resulting in less bias. In addition, by using the novel hybrid framework for robust efficiency analysis (i.e., a Hit-And-Run Monte Carlo simulation), subjectivity is again minimized. Second, analyzing the entire range of admissible values, it would have markedly deteriorated the robustness of the results. For example, such a hypothetical country would be allowed to attain the best and the worst performances on a given indicator. Due to the application of a ratio-based efficiency model, such a unit would attain the extreme comprehensive performances (e.g., the first and the last ranks depending on the precise performance adopted in a given scenario) only because of treating the missing values in this particular way.

The next step in the preparation of the data for further analysis is the normalization. In fact, the algorithm used in this research is a weighted sum over another weighted sum (called the ratio-based efficiency measure, see formula 1), which makes normalization a necessary step to render the different measurement units comparable [79]. The normalization method adopted in the present case study is the target one, which consists in dividing all the indicators by their maximum value, because it conserves the ratios. Hence, each indicator has a maximum value of 1 and a minimum value in the range [0, 1]. Additionally, the ratios are conserved (see Table S2 in the electronic supplementary information (ESI)).

Finally, DEA requires the indicators to be classified into either being inputs or outputs (see formula 1). As the present case study represents a general benchmarking problem, where DEA is employed for decision-making, the inputs are the indicators with a negative preference order (i.e., to be minimized) and the outputs are the ones with a positive preference order (i.e., to be maximized) [78]. Hence, the classification of the indicators is straightforward and univocal. This results in a total of 4 inputs (i_1 , i_2 , i_6 and i_{11}) and 8 outputs (i_3 , i_4 , i_5 , i_7 , i_8 , i_9 , i_{10} and i_{12}) (see Table 2).

Table 2. Performance matrix with selected countries and 12 indicators. The table including all 140 countries is available in Table S1 in the electronic supplementary information (ESI). For the preference order of the values, an upward pointing arrow indicates better performance for higher values (positive preference order), whereas a downward pointing arrow indicates better performance for lower values (negative preference order).

0-Country	1-SAIDI	2-Accident Risks	3-Control of Corruption	4-Political Stability and Absence of Violence/Terrorism	5-Electricity Mix Diversity	6-Electricity Import Dependence	7-Equivalent Availability Factor	8-GDP per Capita	9-Insurance Penetration	10-Government Effectiveness	11-Average Outage Time	12-Ease of Doing Business
Measurement Unit	Hours per Customer per Year	Fatalities/r GWey	Percentile Rank	Percentile Rank	Normalized Shannon Index	Ratio Consumption/Production	%	2010 USD per Capita	% of GDP	Percentile Rank	Hours	Distance to Frontier (100 = Best, 0 = Worst)
Preference Order of the Values	↓	↓	↑	↑	↑	↓	↑	↑	↑	↑	↓	↑
Albania	111.8	7.03	38.46	58.10	0.00	1.02	39%	4543	0.90%	53.85	2.15	66.06
Cambodia	34.2	3.53	12.02	49.52	0.39	1.16	64%	1021	0.30%	25.00	1.40	52.34
Canada	1.0	0.02	95.19	95.24	0.60	0.81	56%	50,108	4.23%	94.71	0.71	80.34
China	1.3	9.65	48.56	26.19	0.42	0.95	74%	6498	1.63%	68.27	3.25	63.43
Congo, Dem. Rep.	92.6	7.01	8.17	4.76	0.01	0.81	39%	384	0.30%	2.88	2.03	34.54
Denmark	0.4	0.04	98.08	76.67	0.57	1.14	54%	60,037	2.69%	97.60	0.80	83.91
Eritrea	92.6	0.95	7.69	19.05	0.01	0.86	85%	528	0.40%	4.81	2.03	26.16
Finland	0.2	0.03	99.52	87.14	0.71	1.20	73%	45,208	2.18%	96.15	0.50	80.34
France	0.1	0.01	88.94	51.43	0.39	0.82	80%	41,768	3.09%	88.46	1.00	75.19
Germany	0.3	0.07	93.27	68.57	0.74	0.89	71%	45,252	3.36%	93.75	1.50	78.52
Haiti	92.6	1.44	10.10	22.38	0.12	0.41	81%	728	1.00%	0.96	2.03	38.63
Iceland	0.5	0.00	95.67	95.71	0.25	0.97	53%	45,939	2.20%	90.87	1.00	78.33
Iraq	2352.0	0.97	4.81	2.86	0.31	0.64	83%	5120	1.63%	9.62	2.33	44.56
Italy	0.7	0.05	57.69	58.57	0.76	1.09	68%	33,912	2.06%	69.23	0.28	71.16
Japan	0.4	0.08	91.35	89.05	0.65	0.96	78%	47,142	2.55%	95.19	4.00	75.36
Kenya	188.5	2.88	13.94	9.52	0.46	0.81	69%	1134	1.88%	43.27	11.42	54.19
Libya	1883.4	0.50	0.96	3.33	0.30	0.28	85%	5447	0.40%	1.92	3.11	32.84
Luxembourg	0.2	0.02	97.12	98.10	0.49	2.97	54%	108,965	1.79%	93.27	1.00	68.77
Myanmar	92.6	4.20	20.67	10.48	0.33	0.84	58%	1643	0.10%	10.10	2.03	38.68
Nepal	92.6	7.02	32.21	14.29	0.01	1.12	39%	690	1.63%	12.98	2.03	59.99
Netherlands	0.3	0.08	94.71	80.48	0.57	1.03	80%	51,285	8.35%	97.12	1.00	75.21

Table 2. Cont.

0-Country	1-SAIDI	2-Accident Risks	3-Control of Corruption	4-Political Stability and Absence of Violence/Terrorism	5-Electricity Mix Diversity	6-Electricity Import Dependence	7-Equivalent Availability Factor	8-GDP per Capita	9-Insurance Penetration	10-Government Effectiveness	11-Average Outage Time	12-Ease of Doing Business
Measurement Unit	Hours per Customer per Year	Fatalities/r GWey	Percentile Rank	Percentile Rank	Normalized Shannon Index	Ratio Consumption/Production	%	2010 USD per Capita	% of GDP	Percentile Rank	Hours	Distance to Frontier (100 = Best, 0 = Worst)
Preference Order of the Values	↓	↓	↑	↑	↑	↓	↑	↑	↑	↑	↓	↑
New Zealand	2.4	0.02	100.00	99.05	0.56	0.94	56%	36,236	4.64%	98.56	1.71	86.42
Niger	290.0	0.79	30.77	13.33	0.31	2.07	84%	384	0.70%	31.25	1.50	45.39
Nigeria	2900.5	1.37	12.50	6.19	0.21	0.83	77%	2535	0.20%	16.35	6.38	46.40
North Korea	92.6	5.30	9.13	10.95	0.32	0.84	52%	1068	1.63%	2.40	2.03	62.67
Norway	1.8	0.00	99.04	91.43	0.10	0.84	40%	89,595	2.21%	98.08	0.82	82.49
Paraguay	15.9	7.03	15.87	48.57	0.00	0.20	39%	3822	1.20%	17.31	2.03	59.82
Qatar	0.7	0.12	78.37	84.29	0.00	0.94	85%	74,531	1.50%	77.40	1.75	65.32
Singapore	0.0	0.12	96.63	96.19	0.11	0.98	85%	51,809	1.69%	100.00	0.00	84.60
South Korea	0.0	0.07	66.83	53.81	0.56	0.97	85%	25,021	4.12%	79.81	0.00	83.52
South Sudan	92.6	0.95	0.48	2.38	0.02	0.94	85%	332	1.63%	0.48	2.03	35.70
Spain	0.3	0.05	69.71	55.71	0.84	0.91	66%	30,486	2.75%	85.10	0.50	73.87
Sweden	1.9	0.01	98.56	80.95	0.53	0.82	58%	55,159	1.88%	96.63	1.46	80.23
Switzerland	0.1	0.01	97.60	96.67	0.41	0.92	58%	75,594	4.12%	99.52	1.00	75.80
Syria	92.6	0.52	1.92	0.00	0.31	0.84	84%	919	0.30%	5.29	2.03	41.53
Togo	92.6	5.10	25.48	38.10	0.34	15.06	53%	554	1.10%	11.06	2.03	46.30
Turkmenistan	92.6	0.12	5.77	42.86	0.00	0.73	85%	6937	1.63%	19.23	2.03	62.67
UK	0.4	0.05	93.75	61.43	0.75	0.98	76%	41,196	2.44%	94.23	2.00	82.57
Uruguay	5.6	4.34	89.42	85.24	0.49	0.80	46%	13,950	1.55%	72.60	1.75	61.69
USA	0.6	0.07	89.90	67.14	0.67	0.96	79%	51,593	4.22%	89.90	2.00	82.03
Venezuela	92.6	4.66	4.33	15.71	0.39	0.65	56%	12,793	3.89%	10.58	2.03	35.30
Vietnam	21.4	2.79	41.83	50.00	0.49	0.92	68%	1685	0.74%	55.29	1.98	59.04
Yemen	92.6	0.62	3.37	0.48	0.29	0.74	85%	775	0.20%	3.37	2.03	44.58

3.2. Ratio-Based Efficiency Analysis with the Charnes, Cooper, and Rhodes (CCR) Model

DEA is a method for assessing efficiencies of DMUs, which in the present case study are countries. Considering a set of DMUs ($D = \{DMU_1, \dots, DMU_K\}$, where K is the number of DMUs), the efficiency of $DMU_o \in D$ is calculated as the ratio between virtual output and virtual input, which are quantified as weighted sums of individual outputs and inputs, respectively [27]. The equation is:

$$E_o(v, u) = \frac{\sum_{n \in OUT} u_n y_{no}}{\sum_{m \in IN} v_m x_{mo}} \quad (1)$$

where:

- $E_o(v, u)$ is the efficiency of $DMU_o \in D$;
- x_{mo} is the amount of m -th input consumed by $DMU_o \in D$, $m \in IN$ (by default a set of inputs is defined as $IN = \{1, \dots, M\}$);
- y_{no} is the amount of n -th output produced by $DMU_o \in D$, $n \in OUT$ (by default a set of inputs is defined as $OUT = \{1, \dots, N\}$);
- $v_{IN} = \{v_j : j \in IN\}$: a vector of input weights (by default $v_{IN} = \{v_1, \dots, v_M\}$);
- $u_{OUT} = \{u_j : j \in OUT\}$: a vector of output weights (by default $u_{OUT} = \{u_1, \dots, u_N\}$).

In a standard DEA setting, the aim is to divide the DMUs into efficient and inefficient ones. For this purpose, one has to find for each $DMU_o \in D$ a weight vector that maximizes its efficiency score. Hence, the ratio-based efficiency analysis with the CCR model consists in solving the following primal optimization problem [27]:

$$\begin{aligned} \max E_o^* &= \sum_{n \in OUT} u_n y_{no} \\ \text{subject to : } &\sum_{m \in IN} v_m x_{mo} = 1; \\ &\sum_{n \in OUT} u_n y_{nk} \leq \sum_{m \in IN} v_m x_{mk}; k = 1, \dots, K; \\ &v_m, u_n \geq 0; m \in IN, n \in OUT. \end{aligned} \quad (2)$$

By definition, the DMUs with efficiency score E_o^* equal to 1 are considered as efficient. The rest of the DMUs are inefficient (efficiency scores between 0 and 1 exclusive), because other DMUs or their conical combination achieve higher scores under the same conditions. Note that E_o^* indicates a multiplier that should be applied to all inputs x_{mo} , $m \in IN$ so that $DMU_o \in D$ becomes efficient (e.g., in case $E_o^* = 0.8$, DMU_o would become efficient by decreasing its inputs by 20%).

Moreover, DEA allows to identify benchmarks to be followed and improvement strategies for the inefficient DMUs. These can be determined by solving the following dual optimization problem:

$$\begin{aligned} \min \theta_o \\ \text{subject to : } &\sum_{k=1, \dots, K} \lambda_k x_{mk} \leq \theta_o x_{m0}; m \in IN; \\ &\sum_{k=1, \dots, K} \lambda_k y_{nk} \geq y_{n0}; n \in OUT; \\ &\lambda_k \geq 0, k = 1, \dots, K. \end{aligned} \quad (3)$$

On the one hand, for an efficient $DMU_o \in D$, $\theta_o = 1$ and $\lambda_{k=o} = 1$. On the other hand, for an inefficient $DMU_o \in D$, all DMUs with $\lambda_k > 0$ are contained in the reference set of DMU_o and can be used for constructing a hypothetical reference unit with greater or equal outputs and lower inputs than DMU_o . The differences between the inputs of such a reference unit and DMU_o indicate the improvements of inputs that are expected from DMU_o for attaining the efficiency. Overall, the CCR model allows tackling research questions 1 and 2.

The aforementioned analysis represents an input-oriented perspective. It derives the required reduction of inputs, if any, that would ensure efficiency (i.e., the best ratio between the virtual outputs and inputs for at least one feasible vector of weights associated with these factors), assuming that the outputs of a given DMU remain unchanged. Note that in DEA, it is also possible to conduct an

output-oriented analysis, hence finding the improvements of outputs needed for reaching the efficiency while holding the current amount of inputs (for details, see [27]).

3.3. In-Depth Analysis of Status of Efficiency

Explanation of the efficiency status requires construction of arguments, which can be used to justify its validity and logic. In this section, the task of generating explanations of the outcomes of DEA in view of the following procedures is considered:

- in case $DMU_o \in D$ is efficient, identification of the minimal subsets of indicators that make it efficient (such minimal subsets of inputs and outputs are called efficiency reducts);
- in case $DMU_o \in D$ is inefficient, identification of the smallest subsets of other DMUs that underlie its inefficiency (such minimal subsets of DMUs are denoted as efficiency constructs).

It is to be noted that the efficiency reducts and constructs are new methodological developments, as explained in the following sentences. On the one hand, to determine all efficiency reducts for some efficient $DMU_o \in D$, an additive method is implemented (see Algorithm 1). It consists of a progressive verification if DMU_o is efficient when using different subsets of inputs IN and outputs OUT , while starting with the smallest ones, and eliminating from further consideration the proper supersets (a superset is a set that includes another set. For example, i_1, i_2 and i_3 is a superset of i_1 and i_3 , and i_2 and i_3 .) of these subsets of indicators that already guaranteed the efficiency [80]. For each efficient DMU_o there exists at least one efficiency reduct (in the worst-case scenario, it contains all inputs and outputs).

Algorithm 1. Additive method for identifying all efficiency reducts.

Require: sets of inputs IN and outputs OUT

Ensure: ERs , all efficiency reducts for $DMU_o \in D$

- 1: $IO =$ all subsets containing at least one input from IN and at least one output from OUT ordered with respect to the increasing cardinality
 - 2: **for** each $IO_k \in IO$ **do**
 - 3: Solve equation (2) for $DMU_o \in D$ with inputs and outputs reduced to IO_k to derive an optimal solution $E_o^*(IO_k)$
 - 4: **if** $E_o^*(IO_k) = 1$ **then**
 - 5: $ERs = ERs \cup IO_k$
 - 6: Remove all supersets of IO_k from IO
 - 7: **end if**
 - 8: **end for**
-

On the other hand, to identify an efficiency construct for some inefficient $DMU_o \in D$, the aim is to find a subset of other DMUs that once removed from the analysis would make DMU_o efficient. This can be attained by solving the following mixed-integer linear programming (MILP) problem:

$$\begin{aligned}
 \min f_w &= \sum_{k=1, k \neq 0}^K b_k \\
 \text{subject to : } & \sum_{n \in OUT} u_n y_{no} = \sum_{m \in IN} v_m x_{mo} = 1; \\
 & \sum_{n \in OUT} u_n y_{nk} \leq \sum_{m \in IN} v_m x_{mk} + C b_k (k = 1, \dots, K, k \neq 0); \\
 & b_k \in \{0, 1\} (k = 1, \dots, K, k \neq 0); \\
 & v_m, u_n \geq 0; m \in IN, n \in OUT;
 \end{aligned} \tag{4}$$

where C is a large positive constant. If $b_k = 1$, DMU_k needs to be eliminated to make DMU_o efficient. Hence, the optimal solution of the above MILP (denoted with *; e.g., f_w^*) indicates one of the efficiency constructs $IC_w = \{DMU_k \in D : b_k^* = 1\}$. It is possible to identify other constructs by adding

the constraints that forbid finding again the solutions found in the previous iterations ($w, w - 1, \dots, 1$): $\sum_{DMU_k \in IC_w} b_k \leq f_w^* - 1$ [81].

3.4. Robust Efficiency Analysis

The standard ratio-based efficiency analysis derives the efficiency scores from the best-case scenario for each DMU. As a result, such scores may not be representative, because they may only be achieved for a very limited number of weight vector combinations, which—in addition—are different for each DMU. Therefore, some DMUs may be considered as efficient even though they do not perform particularly well in general. Moreover, the subsets of efficient and inefficient DMUs can be large, and standard DEA methods offer poor arguments to discriminate within these subsets [78]. For this purpose, a variety of measures that reflect how DMUs perform across all feasible vectors of input/output weights are accounted for [28]. These results refer to three different perspectives: cardinal ratings (efficiency scores; answers research question 3), pairwise one-on-one comparisons (preference relations; answers research question 4), and ordinal comparisons of all DMUs (efficiency ranks; answers research question 5). Specifically, linear programming (LP) to derive the following exact outcomes is used:

- maximal E_o^* and minimal $E_{o,*}$ efficiency scores for $DMU_o \in D$ attained in the set of all feasible input/output weights (note that E_o^* corresponds to the score derived from the standard analysis);
- a necessary efficiency preference relation \succeq_E^N , which holds for a pair $(DMU_o, DMU_k) \in D \times D$ in case DMU_o attains efficiency at least as good as DMU_k for all feasible input/output weights;
- the best R_o^* and the worst $R_{o,*}$ efficiency ranks for $DMU_o \in D$, which are derived from the analysis of, respectively, minimal and maximal subsets of DMUs that attain better efficiency than DMU_o for some feasible input/output weights.

The measures convey useful knowledge on the performances of DMUs in the most and least advantageous scenarios as well as for all feasible weight vectors combinations. Nonetheless, the difference between extreme outcomes can, in general, be quite large, whereas the necessary relation can leave many DMUs incomparable. For this reason, a stochastic efficiency analysis based on the Monte Carlo (MC) simulation to estimate the probability of different outcomes is applied [82]. Hence, a large representative set of feasible weight vectors $(v, u)^S$ (with W being the number of samples) is derived, using a dedicated algorithm such as Hit-And-Run [83]. Hit-And-Run samplers have been proven to perform well for problems of larger sizes [83]. Note that each vector from the feasible weight space is assigned equal chances to be hit (uniform distribution). For each $(v, u) \in (v, u)^S$, the efficiency score $E_o(v, u)$ for each $DMU_o \in D$ is computed, which allows us to approximate the following stochastic acceptability indexes:

- an efficiency acceptability interval index $EAI(DMU_o, b_i)$, which is the share of feasible weight vectors for which $DMU_o \in D$ attains an efficiency score in the interval $b_i \subset [0, 1]$ ($i = 1, \dots, B$), where B is the number of subintervals ($\bigcup_{i=1}^B b_i = [0, 1]$; $b_i \cap b_j = \emptyset, i \neq j$). This represents the distribution of scores, providing the performance robustness assessment that answers research question 3;
- an expected (average) efficiency $EE_o = \sum_{(v,u) \in (v,u)^S} E_o(v, u) / W$ for $DMU_o \in D$;
- a pairwise efficiency outranking index $PEOI(DMU_o, DMU_k)$ for $(DMU_o, DMU_k) \in D \times D$, which is the share of feasible weight vectors for which DMU_o is not worse than DMU_k in terms of the efficiency score, i.e., $E_o(v, u) \geq E_k(v, u)$. This answers research question 4 as it indicates how well countries perform in comparison with each other;
- an efficiency rank acceptability index $ERAI(DMU_o, r)$, which is the share of feasible weight vectors for which $DMU_o \in D$ attains r -th rank. This answers research question 5 as it allows to rank the countries;
- an expected (average) rank $ER_o = \sum_{r=1}^K r \cdot ERAI(DMU_o, r)$ for $DMU_o \in D$.

Most importantly, all of the above results indicate how stable the scores, rankings, and relations observed for different DMUs are across all feasible weight vectors including those that are disadvantageous for each DMU. Hence, it is complementary to the CCR model as it provides a more likely and plausible representation compared to the standard efficiency analysis. Moreover, these outcomes offer arguments (e.g., average efficiencies or ranks), which enable univocal rankings. For the computational details, discussion on the properties and detailed interrelations between the outcomes computed with LP and MC simulation, see [28]. Overall, the robust efficiency analysis allows us to explore research questions 3, 4 and 5.

4. Results and Discussion

In this section, the results are discussed according to the research questions formulated in Section 1.

4.1. What Are the Best Performing (i.e., Most Resilient) Countries and What Are the Reasons for This Achievement?

The main interest of decision-makers might be knowing which countries are the best. Therefore, the results of the CCR model are given in Table 3. According to this model, 31 out of the 140 countries are deemed as efficient (having a value of 1). All of these countries have at least some inputs and outputs performing well enough so that with specific weight vectors no other countries do better.

Among the inefficient countries, Togo, Benin, Namibia and the Democratic Republic of Congo (DRC) score the lowest, with maximal CCR efficiencies of 0.040, 0.245, 0.268 and 0.373, respectively. This means that even in the best case, i.e., with the most advantageous weight vector combination, other countries perform significantly better. While the CCR model provides a clear differentiation of scores for inefficient countries, it does not allow us to differentiate the efficient ones, with 31 receiving an efficiency of 1. Therefore, building a univocal ranking is impossible with the CCR model. The differentiation between efficient countries comes in the context of ranking robustness assessment presented in Sections 4.3 and 4.5.

After the efficient countries were identified, in the next step it was analyzed why they are efficient. This can be done by identifying the individual indicators that make the corresponding country efficient, i.e., efficiency reducts, as shown in Table 4. For example, it is possible to find priorities (weight vectors) that make the United States of America (USA) efficient when considering inputs 1 and 6 and outputs 5, 7, 8 and 9 only (the numbers correspond to the indicator numbers). In fact, on the inputs, the USA's SAIDI (i_1) of 0.6 hours per customer per year and its electricity import dependence (i_6) are well-performing. On the outputs, the USA's electricity production mix (i_5) is diverse, its EAF (i_7) is high and both its GDP per capita (i_8) and insurance penetration (i_9) are comparatively high on an international scale. In other words, these indicators represent the strengths of the USA and no other country performs better under such priorities. However, this efficiency might not be obvious to reach as the USA requires, under specific weight vectors, the combination of at least two inputs and at least four outputs.

Table 3. Charnes, Cooper, and Rhodes (CCR) maximum efficiencies of the 140 countries.

Country	CCR	Country	CCR	Country	CCR	Country	CCR
Algeria	1.000	Armenia	0.987	Mauritius	0.858	Brazil	0.700
Australia	1.000	Taiwan	0.983	Cyprus	0.846	Hungary	0.700
Bulgaria	1.000	Israel	0.979	Guatemala	0.843	Pakistan	0.698
Canada	1.000	UK	0.970	Uzbekistan	0.838	Mongolia	0.692
Costa Rica	1.000	Tunisia	0.967	Serbia	0.831	Vietnam	0.683
Czech Republic	1.000	Brunei Darussalam	0.953	Nicaragua	0.825	Peru	0.678
Estonia	1.000	Uruguay	0.953	Turkey	0.821	Latvia	0.671
Finland	1.000	Portugal	0.946	Hong Kong	0.815	El Salvador	0.667
France	1.000	India	0.943	Syria	0.811	Honduras	0.665
Germany	1.000	Venezuela	0.942	Bangladesh	0.807	Botswana	0.654
Haiti	1.000	United Arab Emirates	0.941	Bosnia and Herzegovina	0.806	Montenegro	0.648
Iceland	1.000	Azerbaijan	0.936	Belgium	0.806	Angola	0.625
Italy	1.000	Iran	0.936	Congo, Rep.	0.805	Gabon	0.624
Jamaica	1.000	Oman	0.914	Tanzania	0.803	Suriname	0.599
Kuwait	1.000	Malaysia	0.912	South Africa	0.797	North Korea	0.584
Libya	1.000	Argentina	0.906	Iraq	0.796	Kyrgyzstan	0.576
Luxembourg	1.000	Slovenia	0.905	Dominican Republic	0.791	Malta	0.567
Netherlands	1.000	Slovakia	0.895	Austria	0.791	Zimbabwe	0.564
New Zealand	1.000	Saudi Arabia	0.895	Senegal	0.785	Cambodia	0.552
Norway	1.000	Poland	0.890	Panama	0.779	Myanmar	0.551
Paraguay	1.000	Trinidad and Tobago	0.888	Georgia	0.776	Sudan	0.542
Qatar	1.000	Ukraine	0.886	Bolivia	0.770	Mozambique	0.531
Romania	1.000	Yemen	0.883	Colombia	0.763	Croatia	0.514
Russia	1.000	Denmark	0.880	Kosovo	0.761	Albania	0.502
Singapore	1.000	Morocco	0.877	Ghana	0.756	Zambia	0.488
South Korea	1.000	Indonesia	0.876	Egypt	0.747	Nigeria	0.486
Spain	1.000	Chile	0.874	Eritrea	0.745	Cameroon	0.483
Sweden	1.000	Bahrain	0.867	Greece	0.743	Tajikistan	0.478
Switzerland	1.000	Jordan	0.865	China	0.742	Ethiopia	0.461
Turkmenistan	1.000	Cuba	0.865	Kenya	0.739	Nepal	0.429
USA	1.000	Philippines	0.864	Sri Lanka	0.729	Niger	0.396
Ireland	0.993	Cote d'Ivoire	0.864	Ecuador	0.720	Congo, Dem. Rep.	0.373
Mexico	0.992	Thailand	0.860	Lebanon	0.716	Namibia	0.268
Moldova	0.992	Kazakhstan	0.859	Lithuania	0.707	Benin	0.245
Japan	0.991	Belarus	0.858	South Sudan	0.706	Togo	0.040

Considering another example, Singapore can become efficient under certain weight vectors if inputs 1 and 2 and output 4 are considered. This efficiency redcut corresponds to the SAIDI, where Singapore is the best country in the world as it only experiences less than a minute of electricity supply interruption per customer per year [84], the low severe accidents risks, indicating that Singapore's electricity production mix is safe from the point of view of human fatalities, and its outstanding political stability and absence of violence/terrorism. Hence, Singapore requires only two inputs and one output to reach efficiency, indicating that these are stronger compared to those from the USA.

Furthermore, a rather surprising result is given by Libya. Considering its low indicator performances (e.g., SAIDI of 1883.4 hours per customer per year, high corruption and low political stability, GDP per capita, insurance penetration, government effectiveness and ease of doing business), it is unexpected that Libya still is deemed efficient. However, as it has the best ratios of i_5/i_6 or i_7/i_6 , these subsets of indicators still make it efficient for specific weighting vectors.

Table 4. Minimal subsets of indicators (efficiency reducts) that make the corresponding country efficient. The numbers refer to the indicator number listed in Table 2. The full table for all efficient countries is available in the electronic supplementary information (ESI), Table S3.

Country	Inputs	Outputs	Country	Inputs	Outputs	Country	Inputs	Outputs	
Algeria	1;2;6;11	7		2;6	3		6;11	9	
Australia	1;6	5;7;8		2;6	10		6;11	12	
	6	3;4	Norway (continued)	2;6	12	South Korea (continued)	6;11	3;5	
	6	4;5		2;11	4		6;11	4;5	
	6	5;8		6;11	3		6;11	5;8	
	6	9;10		6;11	10		6;11	7;8	
Canada	1;6	10	Qatar	1;6	7;8	Spain	6	3;5	
	1;6	12		2;6	7;8		6	5;10	
	2;6	10;12		6;11	7;8		6	3	
	2;11	9		1;2	4	Sweden	6	10	
	6;11	10		1;2	8		1;6	5;7;8	
	2;6;11	12		1;6	3		2;6	5	
Estonia	1;2;6	3;7;12		1;6	10		2;6	4;7	
Germany	6	3;5	Singapore	2;11	4		2;6	7;12	
	6	5;8		2;11	8	6	5;8;9		
Haiti	1;6	5;7		6;11	3		1;2	3	
	6;11	7		6;11	4		1;2	4	
Italy	2;11	5		6;11	8		1;2	5	
Jamaica	1;6	7		6;11	10		1;2	7	
Kuwait	1;6	7		1;2	3		1;2	8	
	6;11	7		1;2	5	Switzerland	1;2	9	
	6	5	1;2	7	1;2		10		
Libya	6	7		1;2	9			1;2	12
	2;6	12		1;2	10			2;6	9
Luxembourg	2;11	5;8	South Korea	1;2	12		2;6	4;7	
	1;2;11	8		2;11	3		2;6	7;8	
Netherlands	6	9		2;11	5		2;11	5	
New Zealand	2;6	9;12		2;11	7		6;11	4;7;8;9	
	2	8		2;11	9		6;11	7;8;9;10	
Norway	6	8		2;11	10	Turkmenistan	2;6	7	
	6	4;10		2;11	12	USA	1;6	5;7;8;9	
	1;6	10;12							

Knowing which countries are the best might be the first requirement of a decision-maker, but not all countries perform at the top. Therefore, it is necessary to also look at the inefficient countries, including the reasons why they are inefficient and how they can improve themselves, which leads to Section 4.2 below.

4.2. Why Are Some Countries Inefficient, How Can They Improve Their Scores and What Are Their Benchmarks?

Table 5 shows the projections of inefficient countries onto the efficiency frontier. For example, the projection for Uruguay represents 0.691 of Canada and 0.240 of Sweden (the sum of the shares does not necessarily have to be equal to 1, because conical combinations of existing units are tolerated). This means that the closest virtual country to Uruguay, that is situated on the efficiency frontier, is composed of 0.691 times the indicator values of Canada plus 0.240 times the indicator values of Sweden. The distance to this virtual country represents the closest path for Uruguay to become efficient. In other words, Canada and Sweden are its benchmarks. The higher the share, the closer the original country already is to its benchmark. In this example, Uruguay is already closer to Canada compared to Sweden. The same analysis can be made with other countries. For example, Denmark can become efficient with contributions from six countries and Japan from four.

Table 5. Projection of inefficient countries onto the efficiency frontier. Only the shares of Denmark, Japan and Uruguay are hereby displayed. For these three countries, the shares from other countries are null. The full table for all countries is available in the ESI, Table S4.

Country	Canada	Czech Republic	Germany	Norway	Singapore	South Korea	Spain	Sweden	Switzerland
Denmark	0.251	0.000	0.000	0.008	0.078	0.183	0.145	0.000	0.446
Japan	0.039	0.416	0.427	0.000	0.000	0.000	0.000	0.000	0.225
Uruguay	0.691	0.000	0.000	0.000	0.000	0.000	0.000	0.240	0.000

Furthermore, the necessary improvements for all inputs or outputs that need to be applied to make a certain country efficient are given in Table 6. The values are negative for the inputs, as they need to be decreased, and positive for the outputs because they have to be increased. These necessary improvements have to be applied to all the inputs together or all the outputs together.

Table 6. Necessary improvements to make a country efficient. These improvements need to be achieved on all inputs cumulatively or all outputs cumulatively. The full table for all countries is available in the ESI, Table S5.

Country	i1	i2	i3	i4	i5	i6	i7	i8	i9	i10	i11	i12
Denmark	0.000	-0.001	0.133	0.295	0.091	-0.009	0.327	0.075	0.249	0.202	-0.007	0.175
Japan	0.000	-0.002	0.030	0.008	0.007	-0.001	0.029	0.004	0.098	0.055	-0.182	0.122
Uruguay	-0.002	-0.448	0.044	0.042	0.095	-0.002	0.103	0.333	0.239	0.204	-0.068	0.194

Finally, as DEA is a measure of efficiency in relation to other DMUs, it is also possible to become efficient when the country list is changed. Table 7 represents the efficiency construct to be removed in order to make a given country efficient. In the case of Uruguay, it is inefficient because Canada is included in the analysis. In other words, even under the most favorable conditions for Uruguay, Canada will perform better. Another example is Denmark, which has a high level of electricity supply resilience at an absolute level, but it can do even better as under similar conditions, Switzerland scores higher. Furthermore, for Denmark, Japan and Uruguay, only one country has to be removed to make them efficient. This indicates that these three countries are close to being efficient. On the contrary, for example, the Democratic Republic of Congo (DRC) requires the removal of 119 countries to become efficient (see Table S6). In other words, the DRC is far away from the efficiency frontier. It is important to note that, for policymaking, the efficiency constructs should not be misused. In fact, it would not make sense to adapt the list of countries in order to artificially increase the score of a country of interest. The efficiency constructs should rather be seen as a tool to identify benchmarks towards which inefficient countries should aim for.

The results presented in Sections 4.1 and 4.2 demonstrated that the CCR model successfully identified (1) the benchmarks, (2) the leading indicators for each country from an efficiency perspective,

(3) the distance from the efficiency frontier, (4) the required indicator improvements to make inefficient countries efficient and (5) the list of countries making others inefficient. However, one has to bear in mind that these findings are based on the most advantageous weight vectors for each country, clearly representing best-case scenarios.

Therefore, the analysis is extended using robust efficiency analysis models to address the following questions:

- What is the average or most likely performance of a country? What is the expected distribution of its performance (Section 4.3)? This answers research question 3.
- Is there a univocal ranking of countries (Section 4.4)? This answers research question 4.
- How does each country perform against all others (pairwise comparisons) (Section 4.5)? This answers research question 5.

Table 7. Minimal subset of countries (efficiency constructs) to be removed in order to make the respective country efficient. For Denmark, Japan and Uruguay, the countries not listed all contain zeros. The full table for all countries is available in the ESI, Table S6.

Countries	Canada	Czech Republic	Germany	Norway	Singapore	South Korea	Spain	Sweden	Switzerland
Denmark	0	0	0	0	0	0	0	0	1
Japan	0	0	1	0	0	0	0	0	0
Uruguay	1	0	0	0	0	0	0	0	0

Identifying benchmarks allows us to analyze weaknesses and develop successful policies. However, the results of Sections 4.1 and 4.2 are based on the CCR model, i.e., a best-case scenario, as explained in Section 3.4. Decision-makers might be interested in a more unbiased view of the results in which ranking robustness is assessed. This leads to Section 4.3 below.

4.3. How Robust Is the Performance of the Countries?

Instead of analyzing only the best-case scenario, the Hit-And-Run Monte Carlo-based robust efficiency model accounts for the wide variability of preference models and it calculates a distribution of performance scores of the countries, thus providing a measure of robustness [83,85]. Table 8 shows the results for the maximum and average performance of the scores obtained with 10,000 model runs. Furthermore, the efficiency acceptability interval indices (10 bins of equal size) are displayed. The sum of the efficiency acceptability interval indices per country is equal to 100%. The table shows only a selection of 45 from the 140 countries, ordered by decreasing values of average efficiency.

According to this model, South Korea, Singapore and Canada are the most efficient countries as in 64.1%, 47.8% and 34.7% of the simulations, respectively, their efficiency is situated in [0.9, 1]. As these countries perform well on multiple inputs and multiple outputs, their efficiency score is high for many weight vectors. On the other end of the spectrum, Togo, the DRC and Nigeria have, for more than 98% of the Monte Carlo simulations performed, efficiencies situated in [0, 0.1]. In Togo's or the DRC's case, neither of their indicators is well-performing. Regarding Nigeria, only its EAF (i_7) performs at an average level, which in comparison with other countries, still makes it one of the worst-performing ones.

Table 8. Maximum and average efficiency scores for the robust efficiency analysis model, as well as the efficiency acceptability interval indices. The number after the countries' names corresponds to their ranks computed based on the average efficiency. The full table for all countries is available in the ESI, Table S7.

Country	Simulation-Based Monte Carlo		Simulation-Based Monte Carlo Efficiency Acceptability Interval Indices (in %)									
	Maximum	Average	[0.0–0.1]	[0.1–0.2]	[0.2–0.3]	[0.3–0.4]	[0.4–0.5]	[0.5–0.6]	[0.6–0.7]	[0.7–0.8]	[0.8–0.9]	[0.9–1]
South Korea (1)	1.000	0.911	0.0	0.0	0.1	0.2	0.5	1.1	4.3	10.9	18.8	64.1
Singapore (2)	1.000	0.852	0.0	0.1	0.3	0.6	1.6	4.6	9.4	15.6	19.9	47.8
Canada (3)	1.000	0.668	3.7	6.1	7.1	6.9	8.0	7.5	8.6	7.8	9.6	34.7
Spain (4)	1.000	0.594	2.8	5.2	7.0	7.9	9.1	10.7	15.5	21.0	16.2	4.6
Finland (5)	1.000	0.583	2.2	4.6	6.3	7.5	9.4	12.3	20.2	28.8	7.0	1.7
Norway (6)	1.000	0.576	5.2	8.2	8.9	8.6	9.4	9.3	9.9	11.6	13.4	15.4
Switzerland (7)	1.000	0.571	5.6	8.5	9.0	9.3	8.5	9.2	8.1	9.5	20.3	12.0
Italy (8)	0.919	0.554	1.2	3.5	5.1	7.8	11.3	23.7	31.0	14.1	2.3	0.1
Netherlands (9)	1.000	0.523	5.2	8.9	9.5	9.9	10.5	11.0	13.1	17.9	10.9	3.0
France (10)	1.000	0.516	6.8	10.3	9.9	9.5	10.1	9.1	10.9	15.9	12.9	4.6
Denmark (11)	0.786	0.489	4.8	8.2	9.3	10.6	11.4	13.7	27.8	14.2	0.0	0.0
Iceland (12)	0.984	0.477	6.9	10.5	10.4	10.9	11.1	10.8	15.3	20.1	3.9	0.2
Sweden (14)	1.000	0.471	10.1	12.5	11.4	10.6	9.0	8.8	9.4	13.2	10.3	4.7
Germany (16)	0.975	0.438	10.1	13.1	11.9	12.0	10.2	10.8	12.0	11.9	7.1	0.8
New Zealand (17)	0.915	0.435	11.1	13.1	12.5	11.4	9.3	9.4	11.5	13.7	7.7	0.2
USA (24)	0.949	0.381	13.7	15.1	13.9	11.9	11.1	12.0	11.5	8.1	2.7	0.1
UK (26)	0.916	0.366	14.4	15.4	14.0	12.5	11.4	12.9	11.0	6.9	1.4	0.0
Qatar (35)	0.914	0.318	16.0	17.5	16.7	15.1	13.3	12.2	6.8	1.9	0.3	0.0
Japan (42)	0.886	0.263	26.6	21.6	14.5	12.3	10.0	6.4	5.0	2.8	0.8	0.0
Luxembourg (43)	0.690	0.260	7.2	17.3	40.4	30.0	4.1	0.7	0.2	0.0	0.0	0.0
Algeria (54)	0.769	0.225	16.7	29.3	28.2	16.5	6.8	1.8	0.5	0.1	0.0	0.0
Turkmenistan (75)	0.775	0.143	44.3	30.0	16.2	6.5	1.9	0.8	0.2	0.1	0.0	0.0
United Arab Emirates (77)	0.823	0.141	52.4	23.2	11.8	5.4	3.7	2.3	0.9	0.2	0.0	0.0
Costa Rica (78)	0.863	0.138	57.8	20.4	8.7	5.4	3.0	2.5	1.4	0.6	0.2	0.0
Uruguay (83)	0.749	0.115	60.6	23.0	8.4	4.3	2.0	1.1	0.5	0.1	0.0	0.0
Vietnam (91)	0.548	0.100	63.1	24.3	8.4	3.1	1.0	0.1	0.0	0.0	0.0	0.0
Yemen (98)	0.650	0.083	69.5	22.9	5.7	1.3	0.4	0.1	0.0	0.0	0.0	0.0
Syria (99)	0.600	0.082	69.5	23.5	5.4	1.1	0.4	0.1	0.0	0.0	0.0	0.0
Haiti (102)	0.736	0.077	74.8	18.4	4.4	1.4	0.5	0.3	0.1	0.0	0.0	0.0
Niger (107)	0.299	0.070	77.4	21.7	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Cambodia (110)	0.379	0.068	77.7	17.6	4.2	0.5	0.0	0.0	0.0	0.0	0.0	0.0

When it comes to the distribution of performance, South Korea, Singapore and Canada consistently rank at the top, as they attain efficiencies higher than 0.7 for most of the simulations. This indicates that these three countries clearly outperform the others leading to lower efficiencies for all other countries. In the middle, there are countries that show more balanced distributions of scores. For example, Germany, Sweden, New Zealand and France attained efficiencies in all intervals and the sum of their three highest interval values represent between 37% and 40%. Hence, the efficiency of these countries is highly dependent on the weight vector considered. If more weight is placed on their well-performing indicators, they will score higher, and reciprocally. This aspect is not revealed by the CCR model, as it only retains the single most advantageous weight vector. When analyzing France in more detail, it can be seen that all its indicators are performing well except its political stability and absence of violence/terrorism (i_4), where it ranks 60th, and its relatively low electricity mix diversity (i_5 , 77% of its electricity production comes from nuclear energy). Hence, France mostly performs well, but there exist some disadvantageous weight vectors, resulting in more balanced distribution of scores. Finally, about half of the countries are clustered at the bottom. In fact, for most weight vectors, these countries attain efficiencies situated in $[0, 0.1]$ and only rarely exceed 0.5.

The maximum efficiency column in Table 8 corresponds to the extreme value of the 10,000 Monte Carlo simulations (the convergence of the results was verified by increasing the number of simulations stepwise and the confidence interval of 95% for Stochastic Multicriteria Acceptability Analysis (SMAA) analyses is not exceeded [85]). While the CCR model deemed 31 countries as efficient, the stochastic analysis only identified 13 as efficient, indicating that the Monte Carlo simulation could not find the best scenario for each country based on the chosen number of runs. Both scores would of course be equal, if an infinite amount of simulations would have been performed [28]. Also, as the CCR model is a linear optimization problem, the simulation-based maximum efficiencies will always be lower or equal to the CCR efficiencies. The difference of both scores, along with the efficiency acceptability interval indices, provides an indication of the likelihood of finding weight vectors that result in a high country score. The largest differences between the CCR (best-case) and the simulation-based maximum efficiency were found for Venezuela, Kenya and Luxembourg with differences of 0.430, 0.332 and 0.310, respectively. Among the countries considered efficient by the CCR, the largest differences were found for Luxembourg, Haiti and Algeria (0.310, 0.264, 0.231). This means that, even though these countries can be efficient based on the CCR model, the corresponding weight vectors are limited. This indicates that relying on the CCR model only might be misleading, as some countries are considered efficient even though efficiency is reached for very few and unlikely weight vectors.

Finally, Table 8 also shows the average efficiency, taken as the arithmetic average of all the simulations. Once again, South Korea, Singapore and Canada are the most efficient countries with average efficiency scores of 0.911, 0.852 and 0.668, respectively. On the lower end is again Togo, the DRC and Nigeria. It is important to note that this average efficiency can be used to rank the countries as it is highly unlikely that two or more countries have equal values. Furthermore, a ranking based on the average makes more sense because it is not just driven by the best-case scenario of the CCR model.

Interestingly, even though Libya, among others, reaches maximal efficiency in the CCR model, its average efficiency is extremely low (0.035). In fact, it is lower than numerous countries that do not reach efficiency. Therefore, Libya reaches unitary efficiency only for a very small number of randomly selected weight vectors and hence should not be seen as a country with an overall high electricity supply resilience. As shown in Table 4, only three efficiency reducts make it efficient. These are indicators $\{5, 6\}$, $\{6, 7\}$ and $\{2, 6, 12\}$. In particular, its electricity import dependence is excellent as it produces much more electricity than it consumes and, therefore, can, in the case of shortage, easily cover its own demand. But this particular situation does not mean that Libya's electricity supply is resilient holistically, as 8 out of its 12 indicators perform poorly. On the other end, Denmark is the inefficient country with the highest average efficiency (0.489). In fact, Denmark actually scores high on all of its indicators, except for its EAF (i_7 , 49% of Denmark's electricity production comes from wind energy which has a low EAF) and its insurance penetration (i_9). Nevertheless, for each weight vector, there are

still other countries that perform better. However, based on the average efficiency, one can safely state that Denmark's electricity supply resilience is higher than Libya's. Hence, compared to the CCR model, the stochastic model allows us to distinguish which countries are more robustly deemed efficient than others, even though they all may be considered as efficient in the CCR model. Furthermore, the simulation-based analysis ranks many inefficient countries higher than several efficient ones in the CCR model. Therefore, for policymaking, it is crucial to consider the CCR model in combination with the robust efficiency model, as it gives a more realistic and broader interpretation.

To confirm the robustness of these results, a sensitivity analysis was performed which investigated how the outputs of a model are affected by varying its inputs and which inputs have the highest effect on the results. If the variations in the inputs result in small variations of the outputs, then the model is considered robust. In the present study, the sensitivity analysis was applied on the Monte Carlo simulation by removing each indicator one by one and keeping the others. The Spearman's rank correlation coefficients (ρ) was calculated between the average country efficiencies for the complete indicator set and one indicator removed at a time. It can be seen from Table 9 that all the coefficients are very high and significant at the 0.01 level. The lowest value is for i_2 , where the correlation factor between the vectors of average efficiencies is ca. 0.897. This still represents a very high correlation and the general trend in the ranking is preserved. Thus, it can be concluded that the results are robust.

Table 9. Spearman's rank correlation coefficients (ρ) between vectors of average efficiencies when removing the 12 indicators one by one. All correlation coefficients are significant at the 0.01 level.

Indicator Removed	Correlation
1 (input)	0.998
2 (input)	0.872
3 (output)	0.999
4 (output)	0.998
5 (output)	0.995
6 (input)	0.871
7 (output)	0.996
8 (output)	0.998
9 (output)	0.999
10 (output)	1.000
11 (input)	0.874
12 (output)	0.999

The distribution of performance of the countries presented in this section shows how robust the results are, which is a key interest of decision-makers. Once the robustness is analyzed, it is important for decision-makers to verify the rank of the countries. In fact, a certain score could lead to a high, average or low rank, depending on the performance scores of the other countries. In Section 4.4, it will become clear that there is no one-to-one relationship between scores and ranks.

4.4. What Is the Univocal Ranking of the Countries?

The previous three sections discussed country scores. However, a certain rank can be achieved by different scores and a certain score can result in different ranks. In order to analyze the rank distribution, this section shows the country ranks computed with the simulation-based Monte Carlo analysis (see Table 10). For the optimal weight vectors, the 31 efficient countries obviously have rank 1 as their best. All the inefficient countries rank at best second. Additionally, the expected rank (ER_o) allows us to differentiate all countries, including the efficient ones. As with the average efficiency, the

ER_o can be used to build a univocal country ranking. Therefore, the ER_o once again confirms that the top place is taken by South Korea (ER_o of 3.6). As shown in Section 4.3, the close followers are Canada and Singapore (ER_o of 4.1 and 4.6, respectively). In conformity with previous results, the DRC is the country with the lowest ER_o (137.9).

Table 10. Best and expected ranks for selected countries obtained by the simulation-based Monte Carlo analysis. The three highest efficiency rank acceptability indices (ERAI) are shown with their corresponding ranks. The ERAIs are expressed in %. The number in parenthesis after a country's name is its overall rank according to the expected rank ER_o . The full table for all countries is available in the ESI, Table S9.

Country	Best (R_o^*)	Expected (ER_o)	Rank	ERAI	Rank	ERAI	Rank	ERAI
South Korea (1)	1	3.6	1	46.8	2	25.5	3	5.3
Canada (2)	1	4.1	3	21.5	1	20.9	6	18.1
Singapore (3)	1	4.6	2	42.0	1	25.2	3	4.9
Switzerland (6)	1	8.0	8	14.2	9	12.5	5	8.3
Denmark (11)	2	12.4	10	14.5	11	14.3	9	13.9
USA (28)	1	30.1	30	4.6	29	4.4	26	4.2
UK (32)	2	32.8	35	4.3	30	4.2	30	4.2
Japan (55)	2	53.8	50	2.9	50	2.9	50	2.9
Turkmenistan (77)	1	76.8	66	2.7	77	2.6	70	2.5
Costa Rica (79)	1	79.1	102	2.7	103	2.6	99	2.4
Uruguay (83)	2	83.9	89	3.6	87	3.2	95	3.0
United Arab Emirates (88)	3	87.3	75	2.9	94	2.0	111	1.9
Syria (98)	10	98.9	105	3.5	86	2.8	78	2.7
Niger (99)	44	99.1	89	3.6	87	3.5	87	3.5
Libya (129)	1	126.8	138	12.0	139	11.4	137	10.7
Nigeria (139)	32	135.4	140	31.6	139	19.2	138	8.6
Congo, Dem. Rep. (140)	120	137.9	140	42.1	139	16.5	137	9.0

Furthermore, it is interesting to analyze the distribution of the ranks. The countries with the most concentrated ranks are the top- and bottom-performing ones (e.g., South Korea, Singapore, the DRC, Canada and Nigeria, among others). In fact, these are clustered at the top or at the bottom, hence it is more likely for them to have large ERAIs. On the contrary, the United Arab Emirates (UAE), Costa Rica and Luxembourg show only small ERAIs. In fact, the sum of their three largest ERAIs is only 6.8%, 7.7% and 7.8%, respectively. An extreme case is Libya, which according to the weight vectors considered, can attain every possible rank. However, Libya's ER_o is 126.8, indicating that it most likely does not rank high. It can be seen that Denmark performs much better than Libya as its ER_o is 12.4. Switzerland, the country that makes Denmark inefficient, ranks slightly better at 8.0. Finally, the tie between Niger and Syria is also reflected through their almost identical ER_o (99.1 and 98.9, respectively).

After the analysis of scores and ranks of the countries, it would be valuable for decision-makers to identify close competitors. These would help to identify strengths and weaknesses, therefore supporting the development of realistic targets and appropriate policies. This leads to Section 4.5 below.

4.5. How Well Does One Country Perform in Comparison to Another?

Pairwise efficiency outranking indices (PEOI) are used to compare the performance of countries between each other. It represents the share of simulations where a country performs at least as good or better than another (see Table 11). PEOIs become extremely useful for identifying close competitors. In country rankings, relevant comparisons can be made in smaller peer groups, i.e., with similarly performing or geographically neighboring countries. For example, Syria and Niger are in a tie (50.1%). Hence, these two countries can benchmark their performance and closely monitor each other over time. If a country turns out to outperform its peer, then the policies of the two countries can be analyzed and successful strategies identified.

Table 11. Pairwise efficiency outranking indices (PEOI) for selected pairs of countries. The values indicate the shares (in percentage) of weight vector samples for which a country has an efficiency score not worse than another (i.e., at least as good). Due to the large number of countries in the data set, the complete table is available in the ESI, Table S8. Countries hereby shown include the best- and worst-performing ones, ties and some of the ones discussed in this paper.

Country	Canada	Congo, Dem. Rep.	Denmark	Libya	Niger	Nigeria	Singapore	South Korea	Switzerland	Syria	Togo
Canada	100.0	100.0	100.0	100.0	100.0	100.0	34.1	33.9	97.1	100.0	100.0
Congo, Dem. Rep.	0.0	100.0	0.0	28.0	0.5	46.1	0.0	0.0	0.0	0.1	32.3
Denmark	0.0	100.0	100.0	99.9	100.0	100.0	9.0	5.1	27.9	100.0	100.0
Japan	0.0	100.0	5.6	100.0	97.6	100.0	3.9	3.8	0.3	99.2	100.0
Libya	0.0	72.0	0.1	100.0	4.4	98.2	0.0	0.0	0.0	1.9	62.3
Niger	0.0	99.5	0.0	95.6	100.0	99.6	0.0	0.0	0.0	49.9	99.5
Nigeria	0.0	53.9	0.0	1.8	0.4	100.0	0.0	0.0	0.0	0.0	42.3
Singapore	65.9	100.0	91.0	100.0	100.0	100.0	100.0	41.7	76.1	100.0	100.0
South Korea	66.1	100.0	94.9	100.0	100.0	100.0	58.3	100.0	77.8	100.0	100.0
Switzerland	2.9	100.0	72.1	100.0	100.0	100.0	23.9	22.2	100.0	100.0	100.0
Syria	0.0	99.9	0.0	98.1	50.1	100.0	0.0	0.0	0.0	100.0	97.7
Togo	0.0	67.7	0.0	37.7	0.5	57.7	0.0	0.0	0.0	2.3	100.0
UK	0.0	100.0	11.3	100.0	100.0	100.0	7.9	6.0	0.8	100.0	100.0
Uruguay	0.0	100.0	0.3	92.0	66.8	99.1	0.1	0.3	0.0	66.0	100.0
USA	0.0	100.0	13.5	100.0	100.0	100.0	9.0	7.8	1.2	100.0	100.0

Furthermore, as expected, the top-performing countries (i.e., South Korea, Singapore and Canada) only rarely score lower than other countries. Similarly, on the other end of the spectrum, Togo, the DRC and Nigeria rarely achieve higher efficiencies than the other countries. Furthermore, even though Libya is deemed efficient, it performs better than Denmark in ca. 1‰ of the simulations only (the CCR model deemed Denmark as inefficient). Once again, this confirms that Denmark has a better electricity supply resilience than Libya.

Another interesting case is how Switzerland performs compared to the top three countries. Switzerland wins against Singapore and South Korea in ca. 22%–24% of the simulations, but wins against Canada for only 2.9% of the cases, even though the average efficiency of Canada is lower than that from South Korea and Singapore. This is due to the fact that the Euclidian distance between Switzerland's and Canada's indicator performances is smaller, meaning that Switzerland and Canada have more similar indicator values compared to Singapore and South Korea. Hence, for the weight vectors that are advantageous for Switzerland, Canada almost always performs slightly better, whereas South Korea and Singapore do not.

4.6. How Does the Performance of Countries Vary According to Changes in Selected Indicators?

The previous sections presented results on the performance of countries, their potentials for improvement and their position in a univocal ranking. In this section, DEA is used as a means to make country-specific improvement potential evaluations. In this way, policymakers can detect early warning signals and explore different future pathways, leading to more effective decisions and subsequent implementation of strategies to reach the targets. In the present study, two types of country analyses were applied:

1. Obtain a new country ranking, based on updated indicator values according to specific scenarios (Singapore, Section 4.6.1).
2. Determine the minimal required improvements on the indicators in order to become an efficient country (Japan, Section 4.6.2)

4.6.1. Country Analysis: Singapore's Electricity Supply Resilience

Located in Southeast Asia, the small sovereign city-state and island country of Singapore is often referred to as the Switzerland of Asia (e.g., [86]). It portrays a high standard of living [87], a strong economy [88] and political stability [89]. Furthermore, Singapore is one of the largest and most competitive financial centers in the world [90]. Additionally, it is one of the world's top five oil trading and refining hubs [91] and is home to the world's second busiest container port [92]. Regarding the energy sector, Singapore imports mainly petroleum products, crude oil and natural gas. Furthermore, natural gas is the source for 95.2% of the electricity produced [93]. Currently, the Singaporean government is trying to diversify its energy supply, in order to be able to better cope with supply disruptions and price increases [94]. Its Economic Strategies Committee (ESC) published a report aiming at ensuring energy resilience and sustainable growth [95]. It contains five key strategies: (1) diversifying the energy sources, (2) enhancing infrastructure and systems, (3) increasing energy efficiency, (4) strengthening the green economy and (5) pricing energy right.

Overall Singapore, deemed as efficient with the CCR model (see Table 3), has the second highest average efficiency (see Table 8) and the third best expected rank (see Table 10). Its outstanding infrastructure is ranked the second best in the world [96], which is reflected by its low SAIDI (i_1) of less than a minute per customer per year, and by the fact that there are almost no fatalities related to electricity production (i_2). Furthermore, being among the top performers in controlling the levels of corruption (i_3) and political stability (i_4), Singapore has a stable environment that enables clear policymaking and transparent directives. However, its electricity generation mix diversity (i_5), consisting of 95.2% of natural gas, makes it particularly vulnerable to potential disruptions. Additionally, on its electricity import dependence (i_6), Singapore's electricity grid is currently not strongly connected in the region, even though this might change in the near future as there are growing efforts to establish electricity

interconnections in Southeast Asia, in particular between Singapore, Malaysia and Indonesia [97]. Nowadays, Singapore basically produces what it consumes. If there were a shortage for any reason, only limited amounts of electricity could be imported from its neighbors. Furthermore, its EAF (i_7) is high, which is attributed to the fact that the electricity is mainly produced by natural gas (natural gas has an EAF of 0.85 [74]). Also, Singapore excels with a high GDP per capita (i_8) and an outstanding government effectiveness (i_{10}). This is the result of a series of successful developments undertaken in the past decades [98]. However, its insurance penetration as a percentage of GDP (i_9), a central part for accelerating recovery processes [99], is not expected to increase in the future [75]. Regarding the average outage time (i_{11}), Singapore is the top-performing country having almost no interruptions at all and even if there happens to be one, it usually is so short that it is hardly noticeable by the population. Lastly, Singapore’s ease of doing business (i_{12}) is already the second highest in the world [100].

Based on these premises, two scenario analyses were developed and discussed in the following two sections (Sections 4.6.2 and 4.6.3).

4.6.2. Scenario 1: 8% Solar Photovoltaic Electricity Production

One of the weaknesses of Singapore’s electricity supply is its generation mix diversity (i_5) consisting of 95.2% of natural gas, which makes it particularly vulnerable to potential disruptions. To address this issue, one of the government’s current strategies is to diversify its sources [101], by for example increasing the share of renewables [102]. In particular, by 2030, Singapore has a potential of producing 8% of its electricity by solar photovoltaics (PV). This change affects the following indicators:

- i_2 ; improvement from 0.124 to 0.115 fatalities/GWyr: solar PV has lower fatality rates than natural gas [71].
- i_5 ; improvement from 0.11 to 0.22: replacing natural gas generation by solar PV improves the mix diversity.
- i_7 ; deterioration from 0.85 to 0.79: solar PV has a lower EAF than natural gas [74].

Therefore, as the performance of i_2 and i_5 increases, but that from i_7 decreases, it is not yet clear if this will have a positive effect on Singapore’s electricity supply resilience. However, results show that, in this particular case, it is advisable to pursue this strategy as Singapore’s score effectively improves (see Table 12). In fact, even though Singapore keeps its second position for the average efficiency and the third rank according to the ERAIs, its performance improved, hence reducing the gap to South Korea and Canada.

Overall, this shows that it is not possible to predict a priori if an increase of the share of renewables is good or bad for resilience. It has to be studied on a case-by-case basis. In the present example, it turned out that increasing the share of solar PV generated electricity to 8% is positive for Singapore’s electricity supply resilience.

Table 12. Results for scenario 1. The efficiency acceptability interval indices (EAIIs) and the ERAIs are given in %. Even though Singapore does not get the first position, having 8% of solar photovoltaic (PV) generation improves its electricity supply resilience.

Efficiency Interval	Average Efficiency	[0.0–0.1]	[0.1–0.2]	[0.2–0.3]	[0.3–0.4]	[0.4–0.5]	[0.5–0.6]	[0.6–0.7]	[0.7–0.8]	[0.8–0.9]	[0.9–1]
South Korea	0.905	0.0	0.3	0.0	0.2	0.3	1.1	4.3	11.9	19.7	62.2
Singapore new	0.872	0.1	0.2	0.2	0.3	0.9	2.7	6.2	15.6	22.8	51.0
Canada	0.680	2.6	6.5	6.8	6.5	7.5	8.3	8.1	8.4	8.6	36.7
<i>Singapore original</i>	<i>0.852</i>	<i>0.0</i>	<i>0.1</i>	<i>0.3</i>	<i>0.6</i>	<i>1.6</i>	<i>4.6</i>	<i>9.4</i>	<i>15.6</i>	<i>19.9</i>	<i>47.8</i>

ERA I	ER ₀	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 6	Rank 7	Rank 8	Rank 9	Rank 10
South Korea	3.7	43.2	26.8	6.2	3.1	4.1	2.3	0.9	1.3	0.9	0.7
Canada	4.0	20.4	11.3	22.2	7.3	8.2	17.9	3.6	1.5	4.2	1.3
Singapore new	4.3	27.7	40.6	5.1	4.5	2.7	2.6	0.9	1.4	1.6	1.9
<i>Singapore original</i>	<i>4.6</i>	<i>25.2</i>	<i>42.0</i>	<i>4.9</i>	<i>3.2</i>	<i>2.6</i>	<i>2.5</i>	<i>2.3</i>	<i>1.4</i>	<i>1.6</i>	<i>1.0</i>

4.6.3. Scenario 2: Singapore in 2030

The second scenario for Singapore assumes that in the year 2030 it not only reaches an 8% share of solar PV, but that the following two indicators change as well:

- i_6 ; improvement from 0.98 to 0.92: Singapore’s electricity grid is currently not strongly connected in the region, but its import dependence is expected to decrease as a result of planned interconnections with Malaysia and Indonesia [97] and according to the projected production and consumption in 2030 [103].
- i_8 ; improvement from 51,809 to 67,360 USD/capita: according to predictions, Singapore’s GDP will increase to 67,360 USD/capita in 2030 [104].

With the new values for i_2, i_5, i_6, i_7 and i_8 , Singapore’s expected performance in 2030 will improve even more (see Table 13), and it will overtake South Korea and Canada, i.e., reaching the first position as the most resilient country in the world regarding electricity supply.

Table 13. Results for scenario 2. The EAIIs and the ERAIs are given in %. According to this scenario, Singapore has the best electricity supply resilience in the world.

Efficiency Interval	Average Efficiency	[0.0–0.1]	[0.1–0.2]	[0.2–0.3]	[0.3–0.4]	[0.4–0.5]	[0.5–0.6]	[0.6–0.7]	[0.7–0.8]	[0.8–0.9]	[0.9–1]
Singapore new	0.912	0.0	0.1	0.1	0.2	0.5	1.5	5.5	8.7	17.2	66.2
South Korea	0.878	0.0	0.1	0.1	0.0	0.4	1.7	5.5	17.8	26.3	48.1
Canada	0.662	3.1	7.2	7.2	7.2	7.0	7.6	7.2	10.3	11.3	31.9
<i>Singapore original</i>	<i>0.852</i>	<i>0.0</i>	<i>0.1</i>	<i>0.3</i>	<i>0.6</i>	<i>1.6</i>	<i>4.6</i>	<i>9.4</i>	<i>15.6</i>	<i>19.9</i>	<i>47.8</i>

ERAI	ER _o	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 6	Rank 7	Rank 8	Rank 9	Rank 10
Singapore new	3.2	47.2	29.7	3.8	3.3	3.1	1.7	1.1	1.5	1.0	1.2
Canada	3.9	18.8	12.2	24.2	9.1	7.0	17.6	3.5	0.9	3.7	1.2
South Korea	4.1	27.5	39.0	7.1	4.5	3.4	1.7	1.9	1.4	1.4	2.0
<i>Singapore original</i>	<i>4.6</i>	<i>25.2</i>	<i>42.0</i>	<i>4.9</i>	<i>3.2</i>	<i>2.6</i>	<i>2.5</i>	<i>2.3</i>	<i>1.4</i>	<i>1.6</i>	<i>1.0</i>

4.6.4. Country Analysis: Japan’s Electricity Supply Resilience

In the three decades following 1960, Japan’s economy has boomed with average GDP growth rates of up to 12.9% [105] and is currently the third-largest in the world [106]. This lead it to become the fifth best country worldwide, according to a global index from the U.S. News and World Report [107]. Its population enjoys the highest life expectancy in the world [108] proving that the quality of life is high [87]. Japan also ranks particularly high for entrepreneurship (2nd [107]) which portrays its numerous innovations and comparatively high number of patent applications [109]. Regarding the energy sector, Japan has gone through tremendous changes since the 2011 Tōhoku earthquake and tsunami [110]. In fact, due to the Fukushima Daiichi nuclear disaster, the nuclear produced electricity was replaced almost instantaneously by mostly imported oil, gas and coal [68], making Japan more vulnerable to supply disruptions. As this option comes with risks and drawbacks (e.g., import dependence, environmental concerns [111], financial burden [112]), the Japanese government published a revised version of its Strategic Energy Plan (SEP) in 2014 [113]. Its goals are to ensure a stable supply, enhance economic efficiency on the premise of safety and pursue environmental suitability.

Based on these premises, Japan’s low SAIDI (i_1) and low fatality rates related to electricity production (i_2) reflect the overall high quality of its infrastructure [96]. Furthermore, having low levels of corruption (i_3) [114] and high political stability (i_4), Japan has a stable environment that enables clear policymaking and transparent directives. Even though Japan’s performance on the electricity mix diversity (i_5) is far from alarming, it shows slightly lower scores compared to some years ago, as its mix diversity has since the Fukushima Daiichi nuclear disaster increasingly been dependent on imported fossil fuels. Regarding the electricity import dependence (i_6), due to its geographical location, Japan’s electricity grid is currently isolated [115], which means that the production simply follows the consumption pattern. Recently, an interconnected Northeast Asia (NEA) grid has received increasing attention. However, modest economic benefits are a major problem for its implementation [115,116].

Additionally, if this plan were to be realized, it is likely that Japan would overall import electricity which would further increase its dependence on its neighbors. Furthermore, its EAF (i_7) is high, which is attributed to the fact that the electricity is mainly produced by fossil fuels (Fossil fuels have EAFs of 0.85 [74]). Also, Japan shows high GDP per capita (i_8) and government effectiveness (i_{10}). Its insurance coverage (i_9) is expected to grow [75], as a result of increased awareness of potential losses due to frequent recent natural catastrophes [117]. In fact, Japan is the fourth most exposed country in the world [118]. This is probably also the reason why its average outage time (i_{11}) is long (4 hours), even though there is on average only one interruption per citizen every tenth year [100]. Lastly, Japan currently ranks 34th in the ease of doing business ranking [100].

Overall, Japan performs well on most of its indicators. Its average efficiency is 0.263 (42nd, see Table S7) and it ranks 55th according to its expected rank. However, it still does not reach efficiency (CCR score of 0.991, see Table 3). Based on these premises, Section 4.6.5. describes Japan's scenario analysis and its results.

4.6.5. Scenario 3: Required Electricity Generation Portfolio to Make Japan Efficient

Unlike the two scenarios for Singapore, where the original calculations were made with an updated data set, the scenario for Japan is an optimization scenario, where the aim was to find the minimal improvement that makes Japan efficient. The first step was to determine which indicators should be varied. The second step consisted in calculating the minimum required performance changes for these indicators in order to make Japan efficient. The chosen scenario investigated if by varying only the electricity generation portfolio Japan can become efficient. As a consequence, the minimum improvements for indicators i_2 , i_5 and i_7 were calculated to make Japan efficient (see Table 14). These improvements were obtained by running the CCR model and considering a constant, proportional improvement over the three indicators.

Table 14. Required minimum performance on i_2 , i_5 and i_7 in order to make Japan efficient.

Indicator	i_2 : Severe Accident Risk	i_5 : Electricity Mix Diversity	i_7 : Equivalent Availability Factor
Unit	Fatalities/GWeyr	Normalized Shannon Index	%
Original performance	0.0782	0.6526	78.14
Required performance	0.0765	0.6664	79.80

The third step was to calculate to what electricity generation portfolio these new values correspond. Considering the 10 fuel sources that are currently producing electricity in Japan (see Table 15), there are numerous portfolios that result in the required performance on indicators i_2 , i_5 , i_7 . From an optimization point of view, this is equivalent to an underdetermined system (3 indicators for 10 technologies; fewer equations than unknowns). Hence, for each of the portfolios fulfilling the constraints on the three indicators of Table 14, its Euclidian distance was calculated. Performing the calculations for 10 million randomly selected portfolios, Table 15 shows the 10 closest portfolios to Japan's current one (smallest Euclidian distances). In fact, these represent the ones that require the least amount of change in order to make Japan efficient.

Although the 10 portfolios do not show large differences, they can provide different policy perspectives. For example, if the target is to reduce the amount of fossil fuels (coal, oil and natural gas), then portfolios 3 and 10 are most suitable. These two portfolios come with an increase of biomass (biofuels and waste combined), nuclear, hydropower and geothermal electricity, whereas solar PV and wind decrease. Portfolio 8 is the only one that increases the share of solar PV, which already today is on a sharp rise [68]. However, no portfolio jointly increases the shares of solar PV and wind. Overall, portfolio 8 is closest to the Japanese government's goals [113,119,120], as (1) it decreases the amount of coal and oil, (2) only slightly increases the share of natural gas (currently the cleanest fossil fuel,

especially if used in combination with carbon capture and storage [121]), (3) increases the shares of biomass, solar PV, geothermal and nuclear electricity, but (4) decreases the share of hydropower and wind electricity.

Table 15. 10 closest electricity generation portfolios to Japan's current one (listed in increasing order of Euclidian distance).

Technology Share	Coal	Oil	Natural Gas	Biofuels	Waste	Nuclear	Hydropower	Geothermal	Solar PV	Wind
Original	32.96%	9.85%	39.36%	3.32%	0.66%	0.91%	8.76%	0.25%	3.44%	0.50%
Portfolio 1	32.19%	10.05%	38.84%	4.67%	0.52%	1.38%	8.70%	1.90%	1.75%	0.02%
Portfolio 2	30.71%	9.65%	40.27%	4.15%	1.89%	0.10%	8.16%	2.98%	1.64%	0.46%
Portfolio 3	32.34%	9.02%	37.30%	4.99%	2.75%	1.09%	10.12%	1.45%	0.86%	0.07%
Portfolio 4	32.20%	8.50%	41.23%	3.98%	1.02%	0.76%	6.24%	2.62%	1.78%	1.67%
Portfolio 5	31.14%	9.51%	41.61%	4.69%	0.52%	0.58%	6.25%	2.02%	1.76%	1.91%
Portfolio 6	33.73%	8.43%	38.17%	2.10%	2.13%	3.92%	9.18%	0.73%	0.64%	0.96%
Portfolio 7	32.04%	11.35%	36.59%	4.01%	2.38%	2.33%	10.30%	0.19%	0.29%	0.52%
Portfolio 8	31.72%	7.28%	40.74%	5.02%	2.88%	0.98%	5.88%	1.83%	3.53%	0.13%
Portfolio 9	34.52%	7.01%	38.91%	1.96%	3.31%	1.51%	8.13%	2.34%	1.28%	1.02%
Portfolio 10	31.58%	9.76%	37.66%	1.66%	3.60%	3.38%	11.27%	0.66%	0.31%	0.11%

5. Conclusions and Policy Implications

Starting from a set of 12 indicators, this study uses two DEA models to assess the electricity supply resilience of 140 countries. First, the classical CCR model deemed 31 countries as efficient (score of 1), and hence resilient. For these countries, it is possible to find at least one weight vector under which no other country performs better. To gain insights into these efficient countries, a novel algorithm that allows us to calculate their efficiency reducts was developed. This demonstrated which minimal combinations of indicators can make a country efficient. Furthermore, another novel algorithm was developed to identify the efficiency constructs of each inefficient country. In other words, the minimal subsets of countries that make it inefficient was computed.

Second, a robust efficiency analysis was applied. To the authors' best knowledge, the present study represents the first application of such an analysis to a country ranking. A distribution of efficiency scores for each country is calculated, which provides information about ranking stability as it depicts the likelihood of a country scoring in a certain performance bin. Additionally, it allows calculating both the average efficiency and the expected rank of a country that can be used to establish a univocal country ranking. The robustness analysis also allows computing the pairwise efficiency outranking indices.

Finally, scenario analyses for Singapore and Japan were carried out. For Singapore, the analysis consisted in verifying if its current energy policies lead to an even higher resilience, even though Singapore is already efficient according to the CCR model. Results showed that increasing electricity production from solar PV is beneficial for Singapore's electricity supply resilience. In contrast, as Japan is an inefficient country, an optimization problem was solved to determine the minimal required improvement on selected indicators in order to make it efficient. From a policymaking perspective, this is equivalent to finding the optimal way to allocate resources in order to increase its rank. By strictly considering technologies that are already producing electricity, results showed that it is possible to reach efficiency by only slightly changing the production shares.

Overall, this study showed that combining the CCR model, including its efficiency reducts and constructs, with the robust efficiency analysis provides a holistic assessment methodology that can be applied to the present electricity supply resilience assessment of 140 countries, but also similar problems in other domains to support robust decision-making by stakeholders. In fact, even though the CCR model is the most widely used, its results are limited and can be misleading. While the CCR model provides a clear differentiation of scores for inefficient countries, it does not differentiate between the efficient ones. Therefore, building a univocal ranking is impossible. Furthermore, the CCR model provides a best-case scenario, as it computes the most advantageous weight vector for

each country separately. As a result, such scores may not be representative, because they might only be achieved for a very limited number of weight vector combinations. Therefore, the authors believe that by using the hereby developed methodology, policymakers would have a broader view of how the alternatives under study perform. Many policies are indeed based on the results of indices obtained by aggregating average values without considering uncertainty or robustness of the results. This might lead to ill-informed decisions. Accounting for uncertainty in input data and problem structure brings a dynamic component to the usual indices that are static.

By considering the CCR and robust efficiency analysis simultaneously, decision-makers can identify close competitors. This provides important learning lessons from comparable countries (so-called benchmarks). Furthermore, this methodology stimulates a multi-disciplinary approach when considering improving the overall performance of a country. In fact, as the indicators are interrelated, multiple specialists should share knowledge in order to tackle the complexity of today's world. Through collaboration between multiple parties, including research institutions, industry and governmental agencies, it would be possible to develop improvement plans and policies to reach predefined targets. The methodology proposed in this paper could provide an interactive discussion platform to lead the decision-making process.

Supplementary Materials: The following are available online at <http://www.mdpi.com/1996-1073/13/7/1535/s1>.

Author Contributions: The conceptualization of the paper was done by P.G., M.C., A.L., M.S., P.B., M.K. and B.S.; P.G., M.C., M.S. and P.B. developed the data set; P.G., M.C., A.L. and M.K. contributed to the model implementation; P.B. and B.S. supervised the progress of the work; P.G. and M.K. wrote the first draft of the manuscript; M.C., A.L., M.S., P.B., M.K. and B.S. reviewed the paper draft; P.G. managed the reviewing and editing process; All authors have read and agreed to the published version of the manuscript.

Funding: The research was conducted at the Future Resilient Systems (FRS) at the Singapore-ETH Centre (SEC), which was established collaboratively between ETH Zürich and Singapore's National Research Foundation (FI 370074011) under its Campus for Research Excellence And Technological Enterprise (CREATE) program. A. Labijak was supported under the Iuventus Plus program (IP2015 029674-0296/IP2/2016/74). M. Kadziński was supported under the SONATA BIS program (DEC-2019/34/E/HS4/00045).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Levin, K.; Cashore, B.; Bernstein, S.; Auld, G. *Worldwide Trends in Energy Use and Efficiency: Key Insights from IEA Indicator Analysis*; International Energy Agency: Paris, France, 2008.
2. Tang, Y.; Bu, G.; Yi, J. Analysis and lessons of the blackout in Indian power grid on 30 and 31 July 2012. In *Zhongguo Dianji Gongcheng Xuebao, Proceedings of the Chinese Society of Electrical Engineering, Beijing, China, 30–31 June 2012*; Chinese Society for Electrical Engineering: Beijing, China, 2012.
3. European Network of Transmission System Operators for Electricity. *Report on Blackout in Turkey on 31st March 2015*; ENTSO-E: Brussels, Belgium, 2015.
4. Ji, C.; Wei, Y.; Mei, H.; Calzada, J.; Carey, M.; Church, S.; Hayes, T.; Nugent, B.; Stella, G.; Wallace, M.; et al. Large-scale data analysis of power grid resilience across multiple US service regions. *Nat. Energy* **2016**, *1*, 1–8. [[CrossRef](#)]
5. Kröger, W. Securing the Operation of Socially Critical Systems from an Engineering Perspective: New Challenges, Enhanced Tools and Novel Concepts. *Eur. J. Secur. Res.* **2017**, *2*, 39–55. [[CrossRef](#)]
6. Wender, B.A.; Morgan, M.G.; Holmes, K.J. Enhancing the Resilience of Electricity Systems. *Engineering* **2017**, *3*, 580–582. [[CrossRef](#)]
7. Gasser, P.; Lustenberger, P.; Cinelli, M.; Kim, W.; Spada, M.; Burgherr, P.; Hirschberg, S.; Stojadinovic, B.; Sun, T.Y. A review on resilience assessment of energy systems. *Sustain. Resilient Infrastruct.* **2019**, 1–27. [[CrossRef](#)]
8. Jovanović, A.; Klimek, P.; Choudhary, A.; Schmid, N.; Linkov, I.; Øien, K.; Vollmer, M.; Sanne, J.; Andersson, S.; Székely, Z. Analysis of Existing Assessment Resilience Approaches, Indicators and Data Sources.

2016. Available online: <https://www.ivl.se/download/18.4a88670a1596305e782de/1484131257184/E002.pdf> (accessed on 6 May 2019).
9. Organisation for Economic Co-operation and Development. International Comparison Program. 2005. Available online: <https://stats.oecd.org/glossary/detail.asp?ID=6280> (accessed on 30 August 2018).
 10. Sovacool, B.K.; Mukherjee, I. Conceptualizing and measuring energy security: A synthesized approach. *Energy* **2011**, *36*, 8. [[CrossRef](#)]
 11. Freudenberg, M. *Composite Indicators of Country Performance*; OECD Science, Technology and Industry Working Papers: Paris, France, 2003.
 12. Bandura, R. *A Survey of Composite Indices Measuring Country Performance: 2008 Update*; United Nations Development Programme, Office of Development Studies (UNDP/ODS Working Paper): New York, NY, USA, 2008.
 13. Greco, S.; Ehrgott, M.; Figueira, J.R. *Multiple Criteria Decision Analysis: State of the Art Surveys*, 2nd ed.; State of the Art Surveys, International Series in Operations Research & Management Science; Springer Science & Business Media: Berlin, Germany, 2006; Volumes 1–2.
 14. Hughes, L.; Shupe, D. *Creating Energy Security Indexes with Decision Matrices and Quantitative Criteria*; Energy Research Group: Halifax, NS, Canada, 2010.
 15. Wu, G.; Liu, L.-C.; Han, Z.-Y.; Wei, Y.-M. Climate protection and China's energy security: Win-win or tradeoff. *Appl. Energy* **2012**, *97*, 157–163. [[CrossRef](#)]
 16. Kaya, T.; Kahraman, C. Multicriteria decision making in energy planning using a modified fuzzy TOPSIS methodology. *Expert Syst. Appl.* **2011**, *38*, 6577–6585. [[CrossRef](#)]
 17. Antanasijević, D.; Pocajt, V.; Ristić, M.; Perić-Grujić, A. A differential multi-criteria analysis for the assessment of sustainability performance of European countries: Beyond country ranking. *J. Clean. Prod.* **2017**, *165*, 213–220. [[CrossRef](#)]
 18. Chung, E.-S.; Lee, K.S. Prioritization of water management for sustainability using hydrologic simulation model and multicriteria decision making techniques. *J. Environ. Manag.* **2009**, *90*, 1502–1511. [[CrossRef](#)]
 19. Valdés, J. Arbitrariness in Multidimensional Energy Security Indicators. *Ecol. Econ.* **2018**, *145*, 263–273. [[CrossRef](#)]
 20. Pohekar, S.; Ramachandran, M. Application of multi-criteria decision making to sustainable energy planning—A review. *Renew. Sustain. Energy Rev.* **2004**, *8*, 365–381. [[CrossRef](#)]
 21. Thies, C.; Kieckhäfer, K.; Spengler, T.S.; Sodhi, M.S. Operations research for sustainability assessment of products: A review. *Eur. J. Oper. Res.* **2018**. [[CrossRef](#)]
 22. Gasser, P. A review on energy security indices to compare country performances. *Energy Policy* **2020**, *139*, 111339. [[CrossRef](#)]
 23. Gasser, P.; Lustenberger, P.; Sun, T.; Kim, W.; Spada, M.; Burgherr, P.; Hirschberg, S.; Stojadinović, B. *Security of Electricity Supply Indicators in a Resilience Context, Proceedings of the European Safety and Reliability Conference, Portorož, Slovenia, 18–22 June 2017*; Taylor & Francis: Portorož, Slovenia, 2017.
 24. Gasser, P.; Suter, J.; Cinelli, M.; Spada, M.; Burgherr, P.; Hirschberg, S.; Kadziński, M.; Stojadinović, B. Comprehensive resilience assessment of electricity supply security for 140 countries. *Ecol. Indic.* **2020**, *110*. [[CrossRef](#)]
 25. Cooper, W.W.; Seiford, L.M.; Zhu, J. Data envelopment analysis. In *Handbook on Data Envelopment Analysis*; Springer: Heidelberg, Germany, 2004; pp. 1–9.
 26. El Gibari, S.; Gómez, T.; Ruiz, F. Building composite indicators using multicriteria methods: A review. *J. Bus. Econ.* **2018**. [[CrossRef](#)]
 27. Charnes, A.; Cooper, W.W.; Rhodes, E. Measuring the efficiency of decision making units. *Eur. J. Oper. Res.* **1978**, *2*, 429–444. [[CrossRef](#)]
 28. Kadziński, M.; Labijak, A.; Napieraj, M. Integrated framework for robustness analysis using ratio-based efficiency model with application to evaluation of Polish airports. *Omega* **2017**, *67*, 1–18. [[CrossRef](#)]
 29. Farrell, M.J. The measurement of productive efficiency. *J. R. Stat. Soc. Ser. A* **1957**, *120*, 253–281. [[CrossRef](#)]
 30. Brockhoff, K. Zur Quantifizierung der Produktivität industrieller Forschung durch die Schätzung einer einzelwirtschaftlichen Produktionsfunktion—Erste Ergebnisse. *Jahrb. Nationalökonomie Stat.* **1970**, *184*, 248–276. [[CrossRef](#)]
 31. Liu, J.S.; Lu, L.Y.Y.; Lu, W.-M.; Lin, B.J.Y. Data envelopment analysis 1978–2010: A citation-based literature survey. *Omega* **2013**, *41*, 3–15. [[CrossRef](#)]

32. Zhou, P.; Ang, B.W.; Poh, K.-L. A survey of data envelopment analysis in energy and environmental studies. *Eur. J. Oper. Res.* **2008**, *189*, 1–18. [[CrossRef](#)]
33. Mardani, A.; Zavadskas, E.K.; Streimikiene, D.; Jusoh, A.; Khoshnoudi, M. A comprehensive review of data envelopment analysis (DEA) approach in energy efficiency. *Renew. Sustain. Energy Rev.* **2017**, *70*, 1298–1322. [[CrossRef](#)]
34. Apergis, N.; Aye, G.C.; Barros, C.P.; Gupta, R.; Wanke, P. Energy efficiency of selected OECD countries: A slacks based model with undesirable outputs. *Energy Econ.* **2015**, *51*, 45–53. [[CrossRef](#)]
35. Lozano, S. A joint-inputs Network DEA approach to production and pollution-generating technologies. *Expert Syst. Appl.* **2015**, *42*, 7960–7968. [[CrossRef](#)]
36. Khalili-Damghani, K.; Tavana, M.; Haji-Saami, E. A data envelopment analysis model with interval data and undesirable output for combined cycle power plant performance assessment. *Expert Syst. Appl.* **2015**, *42*, 760–773. [[CrossRef](#)]
37. Wegener, M.; Amin, G.R. Minimizing greenhouse gas emissions using inverse DEA with an application in oil and gas. *Expert Syst. Appl.* **2019**, *122*, 369–375. [[CrossRef](#)]
38. Wang, H. A generalized MCDA–DEA (multi-criterion decision analysis–data envelopment analysis) approach to construct slacks-based composite indicator. *Energy* **2015**, *80*, 114–122. [[CrossRef](#)]
39. Pang, R.-Z.; Deng, Z.-Q.; Hu, J.-L. Clean energy use and total-factor efficiencies: An international comparison. *Renew. Sustain. Energy Rev.* **2015**, *52*, 1158–1171. [[CrossRef](#)]
40. Li, M.; Wang, Q. International environmental efficiency differences and their determinants. *Energy* **2014**, *78*, 411–420. [[CrossRef](#)]
41. Bampatsou, C.; Papadopoulos, S.; Zervas, E. Technical efficiency of economic systems of EU-15 countries based on energy consumption. *Energy Policy* **2013**, *55*, 426–434. [[CrossRef](#)]
42. Cai, B.; Guo, H.; Ma, Z.; Wang, Z.; Dhakal, S.; Cao, L. Benchmarking carbon emissions efficiency in Chinese cities: A comparative study based on high-resolution gridded data. *Appl. Energy* **2019**, *242*, 994–1009. [[CrossRef](#)]
43. Camarero, M.; Castillo, J.; Picazo-Tadeo, A.J.; Tamarit, C. Eco-efficiency and convergence in OECD countries. *Environ. Resour. Econ.* **2013**, *55*, 87–106. [[CrossRef](#)]
44. Chang, M.-C. Energy intensity, target level of energy intensity, and room for improvement in energy intensity: An application to the study of regions in the EU. *Energy Policy* **2014**, *67*, 648–655. [[CrossRef](#)]
45. Cui, Q.; Kuang, H.-B.; Wu, C.-Y.; Li, Y. The changing trend and influencing factors of energy efficiency: The case of nine countries. *Energy* **2014**, *64*, 1026–1034. [[CrossRef](#)]
46. Gómez-Calvet, R.; Conesa, D.; Gómez-Calvet, A.R.; Tortosa-Ausina, E. On the dynamics of eco-efficiency performance in the European Union. *Comput. Oper. Res.* **2016**, *66*, 336–350. [[CrossRef](#)]
47. Halkos, G.; Petrou, K.N. Analysing the Energy Efficiency of EU Member States: The Potential of Energy Recovery from Waste in the Circular Economy. *Energies* **2019**, *12*, 3718. [[CrossRef](#)]
48. Hsieh, J.-C.; Lu, C.-C.; Li, Y.; Chiu, Y.-H.; Xu, Y.-S. Environmental Assessment of European Union Countries. *Energies* **2019**, *12*, 295. [[CrossRef](#)]
49. Hu, J.-L.; Kao, C.-H. Efficient energy-saving targets for APEC economies. *Energy Policy* **2007**, *35*, 373–382. [[CrossRef](#)]
50. Liou, J.-L.; Wu, P.-I. Will economic development enhance the energy use efficiency and CO₂ emission control efficiency? *Expert Syst. Appl.* **2011**, *38*, 12379–12387. [[CrossRef](#)]
51. Ramanathan, R. An analysis of energy consumption and carbon dioxide emissions in countries of the Middle East and North Africa. *Energy* **2005**, *30*, 2831–2842. [[CrossRef](#)]
52. Robaina-Alves, M.; Moutinho, V.; Macedo, P. A new frontier approach to model the eco-efficiency in European countries. *J. Clean. Prod.* **2015**, *103*, 562–573. [[CrossRef](#)]
53. Song, M.-L.; Zhang, L.-L.; Liu, W.; Fisher, R. Bootstrap-DEA analysis of BRICS' energy efficiency based on small sample data. *Appl. Energy* **2013**, *112*, 1049–1055. [[CrossRef](#)]
54. Wang, L.-W.; Le, K.-D.; Nguyen, T.-D. Assessment of the Energy Efficiency Improvement of Twenty-Five Countries: A DEA Approach. *Energies* **2019**, *12*, 1535. [[CrossRef](#)]
55. Zeng, S.; Streimikiene, D.; Baležentis, T. Review of and comparative assessment of energy security in Baltic States. *Renew. Sustain. Energy Rev.* **2017**, *76*, 185–192. [[CrossRef](#)]
56. Zhang, X.-P.; Cheng, X.-M.; Yuan, J.-H.; Gao, X.-J. Total-factor energy efficiency in developing countries. *Energy Policy* **2011**, *39*, 644–650. [[CrossRef](#)]

57. Zhou, P.; Ang, B.W. Linear programming models for measuring economy-wide energy efficiency performance. *Energy Policy* **2008**, *36*, 2911–2916. [CrossRef]
58. Zhou, G.; Chung, W.; Zhang, Y. Measuring energy efficiency performance of China's transport sector: A data envelopment analysis approach. *Expert Syst. Appl.* **2014**, *41*, 709–722. [CrossRef]
59. Zhou, P.; Poh, K.L.; Ang, B.W. Data Envelopment Analysis for Measuring Environmental Performance. In *Handbook of Operations Analytics Using Data Envelopment Analysis*; Springer: Heidelberg, Germany, 2016; pp. 31–49.
60. Zhou, D.Q.; Wu, F.; Zhou, X.; Zhou, P. Output-specific energy efficiency assessment: A data envelopment analysis approach. *Appl. Energy* **2016**, *177*, 117–126. [CrossRef]
61. Kruyt, B.; van Vuuren, D.P.; de Vries, H.J.M.; Groenening, H. Indicators for energy security. *Energy Policy* **2009**, *37*, 2166–2181. [CrossRef]
62. Ang, B.W.; Choong, W.L.; Ng, T.S. Energy security: Definitions, dimensions and indexes. *Renew. Sustain. Energy Rev.* **2015**, *42*, 1077–1093. [CrossRef]
63. Vera, I.; Langlois, L. Energy indicators for sustainable development. *Energy* **2007**, *32*, 875–882. [CrossRef]
64. Patlitzianas, K.D.; Doukas, H.; Kagiannas, A.G.; Psarras, J. Sustainable energy policy indicators: Review and recommendations. *Renew. Energy* **2008**, *33*, 966–973. [CrossRef]
65. Jansen, J.C.; Arkel, W.V.; Boots, M.G. *Designing Indicators of Long-Term Energy Supply Security*; Energy research Centre of the Netherlands ECN: Westerduinweg, The Netherlands, 2004.
66. Molyneaux, L.; Wagner, L.; Froome, C.; Foster, J. Resilience and electricity systems: A comparative analysis. *Energy Policy* **2012**, *47*, 188–201. [CrossRef]
67. Jasiński, D.; Cinelli, M.; Dias, L.C.; Meredith, J.; Kirwan, K. Assessing supply risks for non-fossil mineral resources via multi-criteria decision analysis. *Resour. Policy* **2018**. [CrossRef]
68. International Energy Agency. Statistics. 2015. Available online: <https://www.iea.org/statistics/statisticssearch> (accessed on 8 March 2018).
69. Gasser, P.; Suter, J.; Cinelli, M.; Lustenberger, P.; Kim, W.; Spada, M.; Burgherr, P.; Hirschberg, S.; Stojadinović, B. Development of an Indicator Set for Resilience Quantification of Electricity Supply. In Proceedings of the Society for Risk Analysis 2017 Annual Meeting, Arlington, VA, USA, 10–14 December 2017.
70. World Bank. Distance to Frontier and Ease of Doing Business Ranking. 2017. Available online: <http://www.doingbusiness.org/~{}media/WBG/DoingBusiness/Documents/Annual-Reports/English/DB17-Chapters/DB17-DTF-and-DBRankings.pdf> (accessed on 6 May 2019).
71. Burgherr, P.; Hirschberg, S. Comparative risk assessment of severe accidents in the energy sector. *Energy Policy* **2014**, *74* (Suppl. 1), S45–S56. [CrossRef]
72. International Renewable Energy Agency. Renewable Energy Statistics. 2017. Available online: <https://www.irena.org/publications/2017/Jul/Renewable-Energy-Statistics-2017> (accessed on 6 May 2019).
73. World Bank. World Governance Indicators. Available online: <http://databank.worldbank.org/data/reports.aspx?source=worldwide-governance-indicators> (accessed on 29 May 2017).
74. Volkart, K.; Bauer, C.; Burgherr, P.; Hirschberg, S.; Schenler, W.; Spada, M. Interdisciplinary assessment of renewable, nuclear and fossil power generation with and without carbon capture and storage in view of the new Swiss energy policy. *Int. J. Greenh. Gas Control* **2016**, *54*, 1–14. [CrossRef]
75. Swiss, R. Sigma Explore—Catastrophe and Insurance Market Data. Available online: <http://www.sigma-explorer.com> (accessed on 8 March 2018).
76. Joint Research Centre of the European Commission. *Handbook on Constructing Composite Indicators: Methodology and User Guide*; OECD publishing: Paris, France, 2008.
77. Meyer, P.; Olteanu, A.-L. Handling imprecise and missing evaluations in multi-criteria majority-rule sorting. *Comput. Oper. Res.* **2019**, *110*, 135–147. [CrossRef]
78. Cook, W.D.; Tone, K.; Zhu, J. Data envelopment analysis: Prior to choosing a model. *Omega* **2014**, *44*, 1–4. [CrossRef]
79. Sarkis, J. Preparing your data for DEA. In *Modeling Data Irregularities and Structural Complexities in Data Envelopment Analysis*; Springer: Heidelberg, Germany, 2007; pp. 305–320.
80. Kadziński, M.; Corrente, S.; Greco, S.; Słowiński, R. Preferential reducts and constructs in robust multiple criteria ranking and sorting. *OR Spectr.* **2014**, *36*, 1021–1053. [CrossRef]
81. Kadziński, M.; Greco, S.; Słowiński, R. Robust Ordinal Regression for Dominance-based Rough Set Approach to multiple criteria sorting. *Inf. Sci.* **2014**, *283*, 211–228. [CrossRef]

82. Lahdelma, R.; Salminen, P. Stochastic multicriteria acceptability analysis using the data envelopment model. *Eur. J. Oper. Res.* **2006**, *170*, 241–252. [CrossRef]
83. Tervonen, T.; van Valkenhoef, G.; Baştürk, N.; Postmus, D. Hit-and-run enables efficient weight generation for simulation-based multiple criteria decision analysis. *Eur. J. Oper. Res.* **2013**, *224*, 552–559. [CrossRef]
84. Energy Market Authority. Smart Energy Sustainable Future—Energy Market Authority Annual Report 2016/17. Available online: https://www.ema.gov.sg/cmsmedia/Publications_and_Statistics/Publications/EMA%20AR%202016_17.pdf (accessed on 6 May 2019).
85. Tervonen, T.; Lahdelma, R. Implementing stochastic multicriteria acceptability analysis. *Eur. J. Oper. Res.* **2007**, *178*, 500–513. [CrossRef]
86. ValueWalk. Singapore—The Switzerland of Asia. Available online: <http://www.valuewalk.com/2017/05/the-switzerland-of-asia/> (accessed on 12 January 2018).
87. United Nations Development Programme. Human Development Report. 2016. Available online: http://hdr.undp.org/sites/default/files/2016_human_development_report.pdf (accessed on 6 May 2019).
88. The Heritage Foundation. Index of Economic Freedom. 2017. Available online: http://www.heritage.org/index/pdf/2017/book/index_2017.pdf (accessed on 6 May 2019).
89. World Bank. Worldwide Governance Indicators—Political Stability and Absence of Violence/Terrorism. 2015. Available online: <http://databank.worldbank.org/data/reports.aspx?source=Worldwide-Governance-Indicators> (accessed on 23 November 2017).
90. The Z/Yen Group and China Development Institute. The Global Financial Centres Index. 2016. Available online: http://www.longfinance.net/images/gfci/20/GFCI20_26Sep2016.pdf (accessed on 6 May 2019).
91. International Trade Administration. Singapore—Oil and Gas. 2017. Available online: <https://www.export.gov/article?id=Singapore-Oil-and-Gas> (accessed on 12 January 2018).
92. World Shipping Council. Top 50 World Container Ports. 2018. Available online: <http://www.worldshipping.org/about-the-industry/global-trade/top-50-world-container-ports> (accessed on 12 January 2018).
93. Energy Market Authority. Singapore Energy Statistics. 2017. Available online: https://www.ema.gov.sg/cmsmedia/publications_and_statistics/publications/ses17/publication_singapore_energy_statistics_2017.pdf (accessed on 6 May 2019).
94. Ministry of Trade Industry. National Energy Policy Report, Energy for Growth. 2007. Available online: <https://www.mti.gov.sg/-/media/MTI/Resources/Publications/National-Energy-Policy-Report/nepr-2007.pdf> (accessed on 6 May 2019).
95. Economic Strategies Committee. ESC Subcommittee on Ensuring Energy Resilience and Sustainable Growth. 2010. Available online: <https://www.mof.gov.sg/Portals/0/MOF%20For/Businesses/ESC%20Recommendations/Subcommittee%20on%20Ensuring%20Energy%20Resilience%20and%20Sustainable%20Growth.pdf> (accessed on 6 May 2019).
96. FM Global. FM Global Resilience Index. 2018. Available online: <https://www.fmglobal.com/research-and-resources/tools-and-resources/resilienceindex> (accessed on 6 May 2019).
97. Suruhanjaya Tenaga. The National Grid, Strengthening Malaysia’s Framework. 2015. Available online: https://www.st.gov.my/ms/general/add_counter/585/download/read_count (accessed on 6 May 2019).
98. Siddiqui, K. The political economy of development in Singapore. *Res. Appl. Econ.* **2010**, *2*, 1. [CrossRef]
99. Asgary, A.; Ozdemir, A.I.; Gentles, C. Does Insurance Delay or Speed up the Recovery and Reconstruction Process? Evidences from Canada. In *Reconstruction and Recovery in Urban Contexts*; UCL: London, UK, 2015.
100. World Bank. Ease of Doing Business Index. 2016. Available online: <http://data.worldbank.org/indicator/IC.BUS.EASE.XQ> (accessed on 8 March 2018).
101. Economic Strategies Committee. High Skilled People, Innovative Economy, Distinctive Global City. 2010. Available online: <https://www.mof.gov.sg/Portals/0/MOF%20For/Businesses/ESC%20Recommendations/ESC%20Full%20Report.pdf> (accessed on 6 May 2019).
102. National Climate Change Secretariat. Singapore’s Approach to Alternative Energy. 2016. Available online: <https://www.nccs.gov.sg/climate-change-and-singapore/national-circumstances/singapore%27s-approach-to-alternative-energy> (accessed on 12 January 2018).
103. Nian, V. *Long Range Energy Analysis of Singapore’s Electricity Sector Using the TIMES Modeling Framework*; National University of Singapore: Singapore, 2013.

104. Pardee Center for International Futures at the University of Denver. Population and GDP Forecasts. 2018. Available online: http://www.ifs.du.edu/ifs/frm_MainMenu.aspx (accessed on 3 November 2018).
105. World Bank. GDP Growth. 2016. Available online: <https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG> (accessed on 12 June 2018).
106. World Bank. GDP. 2016. Available online: <https://data.worldbank.org/indicator/NY.GDP.MKTP.CD> (accessed on 12 June 2018).
107. U.S. News and World Report. Best Countries. 2018. Available online: <https://media.beam.usnews.com/ce/e7/fdca61cb496da027ab53bef37a24/171110-best-countries-overall-rankings-2018.pdf> (accessed on 6 May 2019).
108. World Health Statistics. Monitoring Health for the Sustainable Development Goals. 2016. Available online: https://www.who.int/gho/publications/world_health_statistics/2016/en/ (accessed on 6 May 2019).
109. World Intellectual Property Organization. World Intellectual Property Indicators. 2015. Available online: http://www.wipo.int/edocs/pubdocs/en/wipo_pub_941_2015.pdf (accessed on 6 May 2019).
110. Norio, O.; Ye, T.; Kajitani, Y.; Shi, P.; Tatano, H. The 2011 eastern Japan great earthquake disaster: Overview and comments. *Int. J. Disaster Risk Sci.* **2012**, *2*, 34–42. [CrossRef]
111. Tsukimori, O. Japan's CO₂ emissions hit record as fossil fuel consumption rises. In *Reuters*; Thomson Reuters Corporation: Toronto, ON, Canada, 2014.
112. Matsuo, Y.; Yamaguchi, Y. *The Rise in Cost of Power Generation in Japan after the Fukushima Daiichi Accident and Its Impact on the Finances of the Electric Power Utilities*; The Institute of Energy Economics: Tokyo, Japan, 2013.
113. Ministry of Economy. Strategic Energy Plan. 2014. Available online: http://www.enecho.meti.go.jp/en/category/others/basic_plan/pdf/4th_strategic_energy_plan.pdf (accessed on 6 May 2019).
114. GAN Integrity Solutions. Business Anti-Corruption Portal. 2015. Available online: <https://www.business-anti-corruption.com> (accessed on 12 June 2018).
115. Asia Pacific Energy Research Centre (APERC). *Electric Power Grid Interconnections in Northeast Asia*; APERC: Singapore, 2015.
116. Otsuki, T.; Mohd Isa, A.B.; Samuelson, R.D. Electric power grid interconnections in Northeast Asia: A quantitative analysis of opportunities and challenges. *Energy Policy* **2016**, *89*, 311–329. [CrossRef]
117. Willis Towers Watson. Asia Insurance Market Report. 2016. Available online: <https://www.willistowerswatson.com/-/media/WTW/PDF/Insights/2017/01/Asia-insurance-market-review-report.pdf> (accessed on 6 May 2019).
118. Welle, T.; Birkmann, J. The World Risk Index—An approach to assess risk and vulnerability on a global scale. *J. Extrem. Events* **2015**, *2*, 1550003. [CrossRef]
119. Mancheva, M. Japan Sets 22–24% Renewables Share Target for 2030. Available online: <https://renewablesnow.com/news/japan-sets-22-24-renewables-share-target-for-2030-479165/> (accessed on 12 June 2018).
120. Ministry of Economy. Japan's Energy Plan. 2016. Available online: http://www.meti.go.jp/english/publications/pdf/EnergyPlan_160614.pdf (accessed on 6 May 2019).
121. Hirschberg, S.; Burgherr, P. *Sustainability Assessment for Energy Technologies*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2015; pp. 1–22.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Publication [P3]

A. Labijak-Kowalska and M. Kadziński. Experimental comparison of results provided by ranking methods in data envelopment analysis. *Expert Systems with Applications*, 173:114739, 2021, DOI: 10.1016/j.eswa.2021.114739.

Number of citations³:

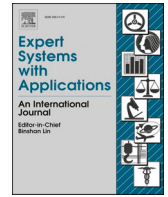
- according to Web of Science: 6
- according to Google Scholar: 11

³as on May 30, 2023



Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Experimental comparison of results provided by ranking methods in Data Envelopment Analysis

Anna Labijak-Kowalska, Miłosz Kadziński *

Institute of Computing Science, Faculty of Computing and Telecommunications, Poznan University of Technology, Piotrowo 2, 60-965 Poznań, Poland

ARTICLE INFO

Keywords:

Data envelopment analysis
Ranking method
Robustness analysis
Monte Carlo simulation
Experimental comparison

ABSTRACT

We consider the problem of ranking Decision Making Units (DMUs) in Data Envelopment Analysis. We illustrate the use of fifteen selected approaches on a numerical example. They represent different categories, including cross- and super-efficiency, multivariate statistics, decision analysis, benchmarking, virtual DMU, and social networks. Moreover, we formalize a new category of ranking methods based on the concept of Robustness Analysis. They exploit a space of feasible input/output weight vectors with the Monte Carlo simulation to derive the expected efficiencies or ranks, or to compute the priorities or net flow scores of DMUs based on the matrix of pairwise efficiency outranking indices. The rankings constructed by all methods are compared on both artificially generated and real-world datasets with different numbers units, inputs and outputs, and performance distributions. The considered datasets represent the most common application areas of the DEA methods, such as finances, education, transportation, healthcare, farming, and the energy industry. The results are quantified in terms of five measures. We indicate that the choice of a method has a significant impact on the ranking, revealing the procedures that offer similar results or differ vastly in terms of the recommended order or the most preferred DMU.

1. Introduction

Data Envelopment Analysis (DEA) is a sub-field of operational research and management science, oriented toward measuring the relative efficiency of Decision Making Units (DMUs) (Cooper et al., 2014). The research on estimating the efficiency frontier in production theory dates back to the work of Farrell (1957). It considers the efficiency defined as the ratio of a single output and a single input. This model was further generalized by Charnes et al. (1978), accounting for a more complex scenario involving multiple inputs and multiple outputs. Specifically, this seminal work refers to an efficiency expressed as the ratio of the virtual output and the virtual input, i.e., weighted sums of outputs and inputs, respectively. The status of efficiency is determined using a linear programming model that compares a given DMU with all other units in the considered set. Such a performance evaluation and measurement is conducted without assigning prior weights and knowing the production function a priori.

The successful applications of DEA can be found in a variety of areas (Emrouznejad and Yang, 2018) such as banking (Thanassoulis, 1999), transportation (Chu et al., 1992), healthcare (Fiallos et al., 2017),

agriculture (Toma et al., 2015), education (Nazarko and Saparuskas, 2014), manufacturing (Bracke et al., 2019), environmental management (Matsumoto et al., 2020), and energy sector (Gasser et al., 2020). Indeed, DEA has the capability of handling complex relations between inputs and outputs of different characters and expressed on various units, while making functional assumptions neither on the considered factors nor on the underlying process (Charnes et al., 1994). For example, in transportation, the efficiency of airports can be evaluated by referring to the inputs capturing the capacities of terminal and apron as well as the catchment area and the outputs corresponding to the numbers of aviation operations and passengers (Kadziński et al., 2017). In manufacturing, the inputs could be raw materials, manpower, floor space, and energy consumption, and the outputs could refer to the numbers of finished goods. In turn, when assessing universities' efficiency, we can consider the inputs in the form of the academic and non-academic staff, operating costs, and area, and the outputs measuring the numbers of students enrolled and completions or research income. In all these contexts, DEA can deliver some objective measures of efficiency, identify the best practice units, and for the under-performing ones – indicate the excess use of inputs or shortfalls in the production of

* Corresponding author.

E-mail addresses: anna.labijak@cs.put.poznan.pl (A. Labijak-Kowalska), milosz.kadziński@cs.put.poznan.pl (M. Kadziński).

<https://doi.org/10.1016/j.eswa.2021.114739>

Received 11 August 2020; Received in revised form 8 December 2020; Accepted 13 February 2021

Available online 21 February 2021

0957-4174/© 2021 Elsevier Ltd. All rights reserved.

outputs.

Traditional DEA models divide the set of considered DMUs into efficient and inefficient ones (Salo and Punkka, 2011). The efficiency of a given DMU is understood so that there is neither any DMU nor any combination of existing DMUs that would attain greater efficiency in the best possible scenario. Thus, all units which are deemed efficient reach an efficiency score of one. In turn, for the inefficient units – this score is lesser than one. From another perspective, efficiency is defined relative to the frontier capturing the observed efficient trade-offs among inputs and outputs for a given set of DMUs. The efficient units lie on the efficient frontier, whereas the relative distance of inefficient units to the frontier is greater than zero (Charnes et al., 1994).

In many decision problems, the binary classification performed with the DEA model is not sufficient because the division into efficient and inefficient units offers too weak discrimination power. When many DMUs are efficient, they cannot be compared, attaining the same maximal efficiency score for different weight vectors. Moreover, when more DMUs are contained in the analyzed dataset, they become too specialized. It is then even more challenging to compare the DMUs, and the classical DEA models fail to provide enriched information about the efficient units. Also, DEA does not perform well for scenarios involving numerous inputs and outputs, as the DMUs tend to become all optimal.

Over the years, numerous solutions have been proposed to increase the discriminatory power of DEA (Kadziński et al., 2017). Specifically, the incorporation of preference information such as weight restrictions, target setting, or value judgments usually leads to fewer efficient DMUs (Podinovski, 2001). Such information may be based on market prices, expert opinion, or preference of a DM (Decision Maker) controlling the units, which are evaluated (for a review, see Joro and Korhonen, 2015). Furthermore, the traditional DEA models are optimistic because each DMU is evaluated in terms of the weights, which are the most advantageous for it. To introduce a joint criterion for the evaluation of all DMUs, techniques based on the common set of weights have been proposed. Such weights form a unique hyperplane as the frontier used to evaluate all DMUs, allowing for their comparison on a common scale. Finally, several ranking approaches have been introduced to impose a complete order on the considered set of DMUs.

Although all ranking methods in DEA aim to order the DMUs from the best to the worst, they are based on different principles. As indicated by the reviews presented by Adler et al. (2002), Aldamak and Zolfaghari (2017) and Hosseinzadeh Lotfi et al. (2013), there already exist a few tens of such methods either implementing a post-analysis to traditional DEA models or offering some dedicated DEA-specific solutions. This paper considers the problem of ranking DMUs with a threefold aim.

First, we illustrate the use of selected ranking methods applicable in the DEA context on the same problem. This serves a didactic purpose as we do not limit ourselves to the presentation of the mathematical model and description of the underlying ranking procedures. In turn, we show the elementary results and explain the underlying steps. This is done in the context of fifteen approaches representative for various categories of ranking methods. These categories include cross-efficiency (Sexton et al., 1986), super-efficiency (Andersen and Petersen, 1993), benchmarking (Lu and Lo, 2009; Jahanshahloo et al., 2007), statistics (Friedman and Sinuany-Stern, 1997; Sinuany-Stern et al., 1994), virtual DMU (Wang and Luo, 2006), social network (Liu and Lu, 2010), and Multiple Criteria Decision Analysis (MCDA) (Sinuany-Stern et al., 2000).

Second, we formalize a new category of ranking methods in DEA, which incorporates Robustness Analysis (RA) (Kadziński et al., 2017). In general, RA accounts for uncertainties observable in the decision problems already at the stage of working out a recommendation (Guo et al., 2019; Kadziński et al., 2020). In the context of DEA, RA refers to the efficiencies attained by the DMUs for all input/output weights or their representative sample. This contrasts with the cross-efficiency technique, which averages the efficiency scores for each DM across a limited subset of weight vectors for which other DMUs attain their maximal efficiency (Sexton et al., 1986). In turn, we incorporate the Monte Carlo

simulation to sample a large set of input/output weight vectors (Lahdelma and Salminen, 2006; Kadziński et al., 2017). For each of them, we compute the efficiency scores and the underlying ranks attained by each DMU.

The proposed category includes two methods that were introduced in the previous works (see Lahdelma and Salminen, 2006; Kadziński et al., 2017) and two novel approaches elaborated in the context of DEA in this paper. Following Lahdelma and Salminen (2006) and Kadziński et al. (2017), in the two RA-based ranking methods we recall, the results for all sampled weight vectors are summarized by the expected efficiency or the expected efficiency rank. These measures can be used to order the DMUs from the best to the worst univocally. With a sufficient number of samples, such robust outcomes accurately approximate the average scores or ranks in the entire space of feasible weights (Tervonen and Lahdelma, 2007). The original methods introduced in this paper consider the matrix of pairwise efficiency outranking indices (Kadziński et al., 2017), which captures the shares of weights for which one DMU attains at least as good score as the other. This matrix is exploited in a twofold way to derive desirability measures for each DMU, given its relative advantage or disadvantage compared with all the remaining units. For this purpose, we incorporate a Net Flow Score procedure (Kadziński and Michalski, 2016) and the eigenvector method (Saaty, 1980).

Third, we conduct the first experimental comparison of ranking methods in DEA regarding the results they provide. For this purpose, we consider two kinds of datasets. On the one hand, we generate artificial sets of DMUs differing in the numbers of units, inputs, and outputs, and performance distributions. On the other hand, we account for ten real-world sets differing for both sizes and application domains. The results are captured by five similarity measures quantifying for all pairs of approaches the agreement between either the entire rankings or the most preferred unit these methods indicate (Kadziński and Michalski, 2016). Such outcomes suggest which ranking procedures offer very similar results and which approaches differ vastly in terms of the recommended order or best choice DMU. The indicated groups of methods can be perceived as suitable substitutes for each other or complementary procedures, offering different perspectives on the ranking of DMUs. This contributes to a better understanding of the comparative strengths and weaknesses of different ranking methods, which have not been well understood until now (Wang, 2020).

The remainder of this paper is organized in the following way. Section 2 introduces the basic concepts of DEA. In Section 3, we describe fifteen ranking methods applicable in the context of DEA and illustrate their use on a numerical example. Section 4 is devoted to an experimental comparison of rankings constructed by different approaches for both artificially generated and real-world datasets. The last section concludes and provides avenues for future research.

2. Notation and basic concepts

The following notation is used in the paper:

- $\mathcal{S} = \{DMU_1, \dots, DMU_K\}$ – a set of considered DMUs, where K is the number of DMUs ($K = |\mathcal{S}|$); each DMU consumes multiple inputs and produces multiple outputs;
- x_m – m -th input, $m \in \{1, \dots, M\}$, where M is the number of inputs;
- y_n – n -th output, $n \in \{1, \dots, N\}$, where N is the number of outputs;
- x_{mo} – an amount or value of m -th input consumed by $DMU_o \in \mathcal{S}$;
- y_{no} – an amount or value of n -th output produced by $DMU_o \in \mathcal{S}$;
- $v = \{v_1, \dots, v_M\}$ – a vector of input weights;
- $u = \{u_1, \dots, u_N\}$ – a vector of output weights;
- $\lambda_{ko}, k = 1, \dots, K$ – the share of k -th DMU in the linear combination when evaluating DMU_o .

2.1. Efficiency analysis

We will use a ratio-based efficiency measure E_o defined as the ratio of the weighted sum of outputs and the weighted sum of inputs of the analyzed $DMU_o \in \mathcal{D}$ (Charnes et al., 1978; Salo and Punkka, 2011):

$$E_o = \frac{\sum_{n=1}^N u_n y_{no}}{\sum_{m=1}^M v_m x_{mo}} \tag{1}$$

To verify the efficiency of DMU_o , Charnes et al. (1978) proposed the following linear programming model, called the CCR model:

$$\begin{aligned} \max \quad & E_o = \sum_{n=1}^N u_n y_{no} \\ \text{subject to :} \quad & \sum_{m=1}^M v_m x_{mo} = 1 \\ & \sum_{n=1}^N u_n y_{nk} \leq \sum_{m=1}^M v_m x_{mk}, \quad k = 1, 2, \dots, K, \\ & u_n, v_m \geq 0, \quad n = 1, 2, \dots, N; m = 1, 2, \dots, M. \end{aligned} \tag{2}$$

It finds the most advantageous set of input/output weights (u_n, v_m) that allow maximizing the efficiency score of DMU_o subject to the assumptions that efficiencies of all DMUs are not greater than one. The obtained solution enables the division of all DMUs into a pair of distinct subsets. The units with an efficiency score equal to one are deemed efficient, whereas those with a score lower than one are inefficient.

Each inefficient unit can be projected onto the efficient frontier. The conical combination of the efficient units, which is the closest to it, indicates the improvements required for becoming efficient. The following model – being the dual form of model (2) – allows finding such a combination for DMU_o :

$$\begin{aligned} \min \quad & \theta \\ \text{subject to :} \quad & \sum_{k=1}^K \lambda_{ko} x_{mk} \leq \theta x_{mo}, \quad m = 1, 2, \dots, M, \\ & \sum_{k=1}^K \lambda_{ko} y_{nk} \geq y_{no}, \quad n = 1, 2, \dots, N, \\ & \lambda_{ko}, \theta \geq 0, \quad k = 1, 2, \dots, K. \end{aligned} \tag{3}$$

The solutions of models (2) and (3) represent the input-oriented perspective. It is focused on the improvement of inputs while keeping the outputs unchanged. It is also possible to formulate the output-oriented counterparts of these models. However, whenever the problem can be analyzed by taking either the input- or the output-oriented perspective, we implement the former without loss of generality.

2.2. Numerical example

Throughout the paper, we will consider an example dataset concerning ten DMUs (A–J) which consume two inputs (i_1 and i_2) and produce two outputs (o_1 and o_2) (see Table 1).

Table 2 presents the efficiency scores obtained with the standard CCR model and the ranks attained by the units according to these scores.

Table 1
A set of ten DMUs considered in the illustrative example.

Unit	i_1	i_2	o_1	o_2
A	347	842	852	356
B	515	136	428	231
C	356	983	12	90
D	851	53	163	626
E	635	554	18	199
F	770	960	285	919
G	73	112	305	54
H	893	847	753	23
I	910	219	197	11
J	687	587	186	502

In particular, there are five efficient DMUs with an efficiency score of one, which ranks them at the very top. The inefficient units are ordered according to their efficiency scores. For example, J, with a score of 0.709, attains the highest rank among the inefficient units, whereas C is ranked last with an efficiency of 0.212.

3. Ranking methods applicable in the context of DEA

In this section, we present fifteen ranking methods applicable in the context of DEA. They represent various categories and impose a complete order on the of DMUs by following some distinct principles. Apart from discussing the underlying mathematical background, we illustrate the use of each method on the same example problem introduced in Section 2. Some of these methods are focused only on ranking the efficient units. In these cases, we assume that the ranks of the inefficient ones remain as imposed by the standard CCR model.

3.1. Cross-efficiency (CE)

Cross-efficiency was introduced by Sexton et al. (1986) to verify the efficiency of all DMUs in different settings. Specifically, K efficiency scores are computed for each DMU, each using the weights forming the most advantageous scenario for some other DMU. In this way, the idea of peer evaluation is implemented, and a common basis in the form of a set of weight vectors is used to compare DMUs. Thus obtained scores are stored in a cross-efficiency matrix, where each cell E_{ij} contains the efficiency attained by DMU_i using the weights optimal for DMU_j . Then, the average efficiency score is computed for each DMU, posing the base for ranking construction:

$$CE_i = \frac{1}{K} \sum_{j=1}^K E_{ij} \tag{4}$$

An important extension of the cross-efficiency model addressing the problem of non-uniqueness of the weights for which a given DMU attains its maximal efficiency was proposed by Doyle and Green (1994). Also, other aggregation operators than the arithmetic mean can be used to derive the final score for each DMU (Green et al., 1996). A detailed discussion on the practical usefulness of this method was provided by Zhu (2014).

The cross-efficiency method requires solving the standard CCR model for each DMU and identifying the weights for which it attained the maximal score. For the considered example, such ten weight vectors are presented in Table 3. The scores achieved by all DMUs for these weights are shown in Table 4. The average scores called cross-efficiencies and the underlying ranks are presented in the last two columns. The efficient units (A, B, D, F, and G) became more comparable than with the standard CCR model. In particular, D proves to be an overall good performer, attaining high scores (greater than 0.6) for the weight vectors being the most advantageous for all DMUs and the maximal score of one under five different scenarios. In turn, A turns out to be a niche performer, being efficient only for its most advantageous scenario, with an efficiency score lower than 0.4 for the three considered settings.

3.2. Super-efficiency (SE)

Super-efficiency was proposed by Andersen and Petersen (1993) by revising the CCR model (Charnes et al., 1978) through eliminating the constraint that limits the efficiency score of the investigated DMU (DMU_o) to values not greater than one, i.e.:

Table 2
Efficiency scores and ranks computed with the CCR model.

	A	B	C	D	E	F	G	H	I	J
CCR score	1.000	1.000	0.212	1.000	0.299	1.000	1.000	0.320	0.286	0.709
Rank	1	1	10	1	8	1	1	7	9	6

Table 3
The inputs/output weights for which each DMU attained its maximal efficiency.

Unit	i_1	i_2	o_1	o_2
A	0.00247	0.00017	0.00027	0.00216
B	0.00052	0.00539	0.00199	0.00064
C	0.00281	0.00000	0.00000	0.00235
D	0.00114	0.00062	0.00000	0.00160
E	0.00107	0.00058	0.00000	0.00150
F	0.00077	0.00042	0.00000	0.00109
G	0.00044	0.00864	0.00328	0.00000
H	0.00006	0.00112	0.00043	0.00000
I	0.00000	0.00457	0.00145	0.00000
J	0.00099	0.00055	0.00020	0.00134

$$\begin{aligned}
 & \max \quad SE_o = \sum_{n=1}^N u_n y_{no} \\
 & \text{subject to :} \quad \sum_{m=1}^M v_m x_{mo} = 1 \\
 & \quad \quad \quad \sum_{n=1}^N u_n y_{nk} \leq \sum_{m=1}^M v_m x_{mk}, \quad k = 1, 2, \dots, K, \quad k \neq o \\
 & \quad \quad \quad u_n, v_m \geq 0, \quad n = 1, 2, \dots, N; \quad m = 1, 2, \dots, M.
 \end{aligned} \tag{5}$$

The optimal solution to the above problem is called super-efficiency. The efficient DMU is allowed to attain a score higher than one. Such a score can be interpreted as the distance of a given DMU from the efficient frontier determined with the CCR model when this DMU cannot participate in its delimitation. In general, super-efficiency quantifies the distance between the efficient frontier and the efficient unit after excluding it. Hence, higher super-efficiency values admit a greater reduction of outputs without losing the status of an efficient unit. Some important revisions of the original super-efficiency approach were proposed, e.g., by Chen (2004), Cook et al. (2009), and Shen et al. (2016).

The super-efficiencies and ranks attained by the efficient units for the illustrative example are presented in Table 5. All these scores are greater than one. Unit D proves to be the most advantageous with super-efficiency equal to 6.954, and A is the worst among the efficient units with a score of 1.065. Nevertheless, all efficient units are guaranteed to be ranked better than inefficient ones. For the latter, super-efficiencies are equal to efficiencies, hence being lesser than one.

3.3. Statistical-based methods

In this subsection, we discuss some ranking approaches based on the multivariate statistical measures to better discriminate between efficient and inefficient DMUs. These approaches aim to derive the sets of

Table 4
Cross-efficiency matrix, cross-efficiency scores, and ranks attained by all DMUs in the illustrative example.

Unit	A	B	C	D	E	F	G	H	I	J	CE	Rank CE
A	1.000	0.408	0.860	0.622	0.622	0.622	0.376	0.376	0.322	0.805	0.601	5
B	0.475	1.000	0.376	0.551	0.551	0.551	1.000	1.000	1.000	0.677	0.718	3
C	0.189	0.015	0.212	0.142	0.142	0.142	0.005	0.005	0.004	0.138	0.099	10
D	0.663	1.000	0.616	1.000	1.000	1.000	0.640	0.640	0.977	1.000	0.854	1
E	0.262	0.049	0.263	0.299	0.299	0.299	0.012	0.012	0.010	0.290	0.179	7
F	1.000	0.208	1.000	1.000	1.000	1.000	0.108	0.108	0.094	1.000	0.652	4
G	1.000	1.000	0.620	0.567	0.567	0.567	1.000	1.000	0.865	1.000	0.819	2
H	0.108	0.301	0.022	0.024	0.024	0.024	0.320	0.320	0.282	0.135	0.156	8
I	0.034	0.241	0.010	0.015	0.015	0.015	0.281	0.281	0.286	0.053	0.123	9
J	0.633	0.197	0.612	0.701	0.701	0.701	0.113	0.113	0.101	0.709	0.458	6

common weights of inputs and outputs, which can be used to evaluate and rank the DMUs. They also admit that some inefficient units are ranked higher than the efficient ones. For some other ranking methods based on statistics and common weights, see Alirezaee and Afsharian (2007), Hatami-Marbini et al. (2015), and Wang et al. (2011). For a hybrid approach combining common weights with Multiple Objective Optimization (MOO), see Carrillo and Jorge (2016).

3.3.1. Canonical correlation analysis (CCA)

The first method based on statistics adapts the Canonical Correlation Analysis (CCA) to the context of DEA (Friedman and Sinuany-Stern, 1997). This approach finds a vector of common weights for all DMUs by incorporating the canonical correlation method for finding the coefficients of linear combinations of inputs or outputs with the maximal correlation degree. Once these coefficients (v_m and u_n) are found, the scaling ratio score, T_o , is computed for each DMU_o as the ratio of the linear combinations of outputs W_o and inputs Z_o , i.e.:

$$T_o = \frac{W_o}{Z_o} = \frac{\sum_{n=1}^N u_n y_{no}}{\sum_{m=1}^M v_m x_{mo}}. \tag{6}$$

3.3.2. Linear discriminant analysis (LDA)

Another DEA ranking method incorporating a statistical analysis was proposed by Sinuany-Stern et al. (1994). It is based on Linear Discriminant Analysis (LDA), aiming to find a vector of input/output weights that separate efficient and inefficient units. Since LDA is designed to detect a linear combination of features, the authors suggested the following formula to convert the efficiency model into a linear function:

$$D_o = \sum_{n=1}^N u_n y_{no} + \sum_{m=1}^M v_m (-x_{mo}). \tag{7}$$

3.3.3. Discriminant analysis of ratios (DR-DEA)

The statistical discriminant analysis (DR-DEA) was introduced by Sinuany-Stern and Friedman (1998). This method avoids the infeasibility issues existing in the methods based on CCA and LDA. The linear combination from LDA is replaced with the ratio of linear combinations of outputs and inputs. This ratio score T_j is defined analogously to the

Table 5
Super-efficiency scores and ranks for the five efficient DMUs in the illustrative example.

	A	B	D	F	G
SE	1.065	1.126	6.954	1.355	2.341
Rank SE	5	4	1	3	2

standard efficiency score:

$$T_j = \frac{\sum_{n=1}^N U_n V_{nj}}{\sum_{m=1}^M V_m X_{mj}} \quad (8)$$

Similar to the LDA method, this approach exploits the division of DMUs into efficient and inefficient units. Specifically, it looks for the set of common weights maximizing the ratio of the between-group variance of T ($SS_B(T)$) and the within-group variance of T ($SS_W(T)$):

$$\begin{aligned} \max \quad & \sum_{n=1}^N s_{no}^y + \sum_{m=1}^M s_{mo}^x \\ \text{subject to:} \quad & \sum_{k=1}^K \lambda_{ko} y_{nk} - y_{no} = s_{no}^y, \quad n = 1, 2, \dots, N \\ & x_{mo} - \sum_{k=1}^K \lambda_{ko} x_{mk} = s_{mo}^x, \quad m = 1, 2, \dots, M \\ & \sum_{k=1}^K \lambda_{ko} = 1 \\ & s_{no}^y \geq 0, s_{mo}^x \geq 0, \lambda_{ko} \geq 0, \quad m = 1, 2, \dots, M; n = 1, 2, \dots, N; k = 1, 2, \dots, K. \end{aligned} \quad (13)$$

$$\begin{aligned} \max \lambda = \frac{SS_B(T)}{SS_W(T)}, \\ SS_B(T) = n_1 (\bar{T}_1 - \bar{T})^2 + n_2 (\bar{T}_2 - \bar{T})^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{T}_1 - \bar{T}_2)^2, \end{aligned} \quad (9)$$

$$SS_W(T) = \sum_{k=1}^{n_1} (T_k - \bar{T}_1)^2 + \sum_{k=n_1+1}^K (T_k - \bar{T}_2)^2,$$

where n_1 and n_2 are, respectively, the numbers of efficient and inefficient DMUs, whereas \bar{T}_1, \bar{T}_2 , and \bar{T} are arithmetic means of ratio scores of the efficient DMUs, inefficient DMUs, and all units, respectively:

$$\bar{T}_1 = \frac{\sum_{k=1}^{n_1} T_k}{n_1}, \quad \bar{T}_2 = \frac{\sum_{k=n_1+1}^K T_k}{n_2}, \quad \text{and} \quad \bar{T} = \frac{\sum_{k=1}^K T_k}{K}. \quad (10)$$

The common sets of weights identified by the three statistical approaches for the illustrative example are presented in Table 6. In the CCA or DR-DEA methods, these weights are incorporated into the standard ratio-based model to obtain each DMU's final score. For example, the score of unit A is computed using DR-DEA as follows:

$$E_A = \frac{1.9245 \cdot 852 + 5.8447 \cdot 356}{8.1711 \cdot 347 + 7.8150 \cdot 842} = \frac{3720.3872}{9415.6017} \approx 0.395. \quad (11)$$

In turn, LDA uses a linear model for aggregation. In the context of unit A, the computation of score can be performed in the following way:

$$E_A = -(0.0050 \cdot 347 + 0.0038 \cdot 842) + 0.0044 \cdot 852 + 0.0066 \cdot 356 \approx 1.20. \quad (12)$$

Table 6
Common input/output weights obtained by the three statistical ranking methods for the illustrative example.

Method	i_1	i_2	o_1	o_2
CCA	0.0023	0.0020	0.0016	0.0031
LDA	0.0050	0.0038	0.0044	0.0066
DR-DEA	8.1711	7.8150	1.9245	5.8447

The scores and ranks of all DMUs considered in the illustrative example, according to the three statistical-based methods, are presented in Table 7. The ranges of these scores are very different. For instance, for CCA, they differ from 1.661 for unit G to 0.106 for unit C. In contrast, for LDA, the highest value of 1.201 is attained by unit A and the least value of -4.854 by unit C. Nevertheless, for all three approaches, the five efficient units are ranked better than the five inefficient ones.

3.4. Benchmark-based methods

The benchmarking approaches quantify the importance of efficient DMUs in terms of their role as a reference for the inefficient units, or, equivalently, their usefulness compared with the other DMUs.

3.4.1. Slack-adjusted efficiency ranking (BSA)

A ranking method for the efficient units based on their importance as benchmarks for the inefficient one has been proposed by Torgersen et al. (1996). Within a two-stage procedure, one first employs an additive model (Charnes et al., 1985) to evaluate the value of slacks of DMUs. All units V with the slacks equal to zero are assumed to be efficient. The underlying model is as follows:

Once the weights are found, the following output-oriented model is applied for each DMU (note that it is possible to replace it with the input-oriented one):

$$\begin{aligned} \max \quad & \frac{1}{E_o} = \phi \\ \text{subject to:} \quad & \sum_{k \in V} \lambda_{ko} y_{nk} - \phi y_{no} = s_{no}^y, \quad n = 1, 2, \dots, N \\ & x_{mo} - \sum_{k \in V} \lambda_{ko} x_{mk} = s_{mo}^x, \quad m = 1, 2, \dots, M \\ & \lambda_{ko} \geq 0, \quad k = 1, 2, \dots, K. \end{aligned} \quad (14)$$

The above model is an output-oriented counterpart of the standard CCR model operating on the linear combinations of DMUs (see model (3)). The variables s_{mo}^x and s_{no}^y are interpreted as the slacks. The slack for a given input/output is the difference between this factor's values for the analyzed DMU and the obtained benchmark (i.e., a linear combination of efficient units).

To construct a ranking of efficient DMUs, for each such a unit DMU_o and n -th output, the potential output value y_{no}^p and the DMU's benchmark measure ρ_o^n have been introduced. The benchmark value is defined as a ratio of the scaled potential increase of n -th output for which a given DMU acts as a referent and all potential increases of this output. The former is defined as a sum, over all units $DMU_k, k = 1, \dots, K$, of the differences between potential output value y_{nk}^p and the actual one y_{nk} multiplied by the benchmark coefficient (λ_{ko}) of DMU_o . Overall, y_{no}^p and ρ_o^n are computed as follows:

Table 7

The scores and ranks of all DMUs in the illustrative example according to the three statistical-based methods (CCA, LDA, and DR-DEA).

	A	B	C	D	E	F	G	H	I	J
CCA	0.985	0.953	0.106	1.053	0.248	0.884	1.661	0.339	0.137	0.664
Rank CCA	3	4	10	2	8	5	1	7	9	6
LDA	1.201	0.347	-4.854	0.440	-3.861	-0.123	0.917	-4.175	-4.410	-1.494
Rank LDA	1	4	10	3	7	5	2	8	9	6
DR-DEA	0.395	0.412	0.052	0.539	0.126	0.429	0.613	0.114	0.048	0.323
Rank DR-DEA	5	4	9	2	7	3	1	8	10	6

Table 8

Slack values and shares in the efficient projection of the inefficient units obtained with the slack-adjusted model (BSA) for the illustrative example.

Unit	s_{i_1}	s_{i_2}	s_{o_1}	s_{o_2}	λ_A	λ_B	λ_D	λ_F	λ_G	θ
C	0.000	114.204	15.911	0.000	0.000	0.000	0.000	0.098	0.000	0.212
E	0.000	0.000	41.520	0.000	0.000	0.000	0.070	0.169	0.000	0.299
H	0.000	0.000	0.000	150.046	0.000	0.256	0.000	0.000	2.110	0.320
I	23.067	0.000	0.000	95.325	0.000	0.460	0.000	0.000	0.000	0.286
J	0.000	0.000	0.000	0.000	0.000	0.000	0.192	0.408	0.126	0.709

$$y_{no}^p = \frac{y_{no}}{E_o} + s_{no}^y, \tag{15}$$

$$\rho_o^n = \frac{\sum_{k=1}^K \lambda_{ko} (y_{nk}^p - y_{nk})}{y_n^p - y_n}, \tag{16}$$

where y_n^p and y_n are sums of, respectively, the potential and actual n -th output values of all DMUs, i.e.:

$$y_n^p = \sum_{k=1}^K y_{nk}^p \quad \text{and} \quad y_n = \sum_{k=1}^K y_{nk}. \tag{17}$$

Having determined all values of ρ_o^n , the efficient DMUs can be ranked according to the values of ρ_o computed as an average of ρ_o^n for DMU_o :

$$\rho_o = \frac{\sum_{n=1}^N \rho_o^n}{N}. \tag{18}$$

When it comes to the illustrative example, we have implemented the input-oriented version of the algorithm for consistency with other ranking methods. In the first stage, the slacks and reference values (λ_{ij}) for the inefficient units are determined using model (14) (see Table 8). These slacks are employed to determine the improved input values. For example, the radial (x_{c2}^R) and improved (x_{c2}^p) values for unit C on i_2 are computed as follows:

$$x_{c2}^R = \theta_c \cdot x_{c2} = 0.212 \cdot 983 = 208.396, \tag{19}$$

$$x_{c2}^p = x_{c2}^R - s_{c2} = 208.396 - 114.204 = 94.192. \tag{20}$$

In the other stage, for each input and DMU, we determine the comprehensive aggregated potential for a decrease of this input for which a given DMU acts as a referent (ρ_m^k). Subsequently, we rank the efficient units according to the comprehensive potential values. Finally, the mean values of these ranks are determined, leading to a complete ranking of efficient DMUs. The values of comprehensive potentials with the underlying ranks and final ranking are presented in Table 9. The comprehensive potentials are by far the greatest for unit G and equal to

Table 9

Values of comprehensive aggregated potentials and ranks for inputs of efficient DMUs along with the average ranks and the final ranks.

Unit	$\rho_{i_1}^k$	$\rho_{i_2}^k$	Average rank	Rank BSA
A	0.000 (5)	0.000 (5)	5	5
B	0.206 (2)	0.112 (2)	2	2
D	0.032 (4)	0.031 (4)	4	4
F	0.085 (3)	0.102 (3)	3	3
G	0.605 (1)	0.633 (1)	1	1

zero for unit A. As a result, these units are ranked first and fifth, respectively.

3.4.2. Ranking based on changing the reference set (BCRS)

The idea underlying another method investigating the role of an efficient DMU as the benchmark for inefficient units derives from measuring the changes of efficiency scores of the inefficient units after removing the examined DMU from the dataset (Jahanshahloo et al., 2007). The more the efficiency frontier approaches the inefficient DMUs, the more efficient is the detached efficient units.

In the first stage, the standard CCR model is applied to divide the units into efficient (J_E) and inefficient (J_I) ones. In the other stage, for each unit (DMU_a) from J_E , we compute an average efficiency score for units from J_I considering a set of DMUs without DMU_a . To determine the efficiency of inefficient unit DMU_b without accounting for DMU_a , the following model needs to be solved:

$$\begin{aligned} \max \quad & E_b^a = \sum_{n=1}^N u_n y_{nb}, \\ \text{subject to:} \quad & \sum_{m=1}^M v_m x_{mb} = 1, \\ & \sum_{n=1}^N u_n y_{nk} \leq \sum_{m=1}^M v_m x_{mk}, \quad k = 1, 2, \dots, K, \quad k \neq a, \\ & u_n, v_m \geq 0, \quad n = 1, 2, \dots, N; m = 1, 2, \dots, M. \end{aligned} \tag{21}$$

Then, an overall score of the efficient DMU, DMU_a , is computed as follows:

$$\Omega_a = \frac{\sum_{b \in J_I} E_b^a}{|J_I|}. \tag{22}$$

The efficiencies E_b^a of inefficient units DMU_b after removing the efficient units DMU_a for the illustrative example are presented in Table 10. The mean values (Ω_a) derived from these efficiencies and the final ranks are provided in the last two columns. G and F attained the

Table 10

Efficiency scores of inefficient DMUs after removal of efficient ones with average values and final ranks.

Unit	E_C^a	E_E^a	E_H^a	E_I^a	E_J^a	Ω_a	Rank BCRS
A	0.212	0.299	0.320	0.286	0.709	0.365	5
B	0.212	0.299	0.325	0.321	0.709	0.373	4
D	0.212	0.344	0.320	0.286	0.815	0.396	3
F	0.246	0.375	0.320	0.286	0.877	0.421	2
G	0.212	0.299	0.627	0.286	0.716	0.428	1

first two ranks mainly because of the high efficiencies attained by H and J, respectively, once these units are removed from the analysis.

3.4.3. Interactive benchmark model (BI)

Another method to rank the units based on the roles of the benchmark they play for other units was proposed by Lu and Lo (2009). In the first stage, this approach fixes each DMU as a benchmark and estimates the efficiency scores for other DMUs. Based on the standard CCR model, the production possibility set is spanned by (x_b, y_b) :

$$P = \{(\bar{x}, \bar{y}) \mid \bar{x} \geq \lambda_b x_b, \bar{y} \leq \lambda_b y_b, \bar{y} \geq 0\}. \tag{23}$$

The subset \bar{P} of the production set P is defined as:

$$\bar{P} = P \cap \left\{ \bar{x} \geq x_o, \bar{y} \leq y_o \right\}, \tag{24}$$

where (x_b, y_b) are vectors of input and output values of benchmark DMU_b , and (x_o, y_o) are vectors of input and output values of examined DMU_o . In the next step, index θ_o^b is computed as the ratio of a weighted distance from (x_o, y_o) to (\bar{x}, \bar{y}) :

$$\theta_o^b = \frac{\frac{1}{M} \sum_{m=1}^M \frac{\bar{x}_m}{x_{mo}}}{\frac{1}{N} \sum_{n=1}^N \frac{\bar{y}_n}{y_{no}}}. \tag{25}$$

Having defined the variables ϕ_m and ψ_n as:

$$\bar{x}_m = x_{mo}(1 + \phi_m) \text{ and } \bar{y}_n = y_{no}(1 - \psi_n), \tag{26}$$

index θ_o^b can be expressed in the following way:

$$\theta_o^b = 1 + \frac{\frac{1}{M} \sum_{m=1}^M \phi_m}{1 - \frac{1}{N} \sum_{n=1}^N \psi_n}. \tag{27}$$

To find the efficiency of DMU_o considering DMU_b as the benchmark, the following problem needs to be solved:

$$\begin{aligned} \min \quad & \theta_o^b = \frac{1 + \frac{1}{M} \sum_{m=1}^M \phi_m}{1 - \frac{1}{N} \sum_{n=1}^N \psi_n} \\ \text{subject to:} \quad & \lambda_{bo} x_{mb} - x_{mo} \phi_m \leq x_{mo}, \quad m = 1, 2, \dots, M \\ & \lambda_{bo} y_{nb} + y_{no} \psi_n \geq y_{no}, \quad n = 1, 2, \dots, N \\ & \phi_m \geq 0, \quad \psi_n \geq 0, \quad \lambda_{bo} \geq 0. \end{aligned} \tag{28}$$

The above model can be transformed into its linear programming counterpart using the Charnes-Cooper transformation (Charnes and Cooper, 1962). For this purpose, we introduce an additional variable t and replace the original variables with the new ones defined as follows: $t_o^b = \theta_o^b$, $\Lambda_{bo} = t\lambda_{bo}$, $\Phi_m = t\phi_m$ and $\Psi_n = t\psi_n$. Then, the resulting model is the following:

$$\begin{aligned} \min \quad & t_o^b = t + \frac{1}{M} \sum_{m=1}^M \Phi_m \\ \text{subject to:} \quad & t - \frac{1}{N} \sum_{n=1}^N \Psi_n = 1, \\ & \Lambda_{bo} x_{mb} - x_{mo} \Phi_m - x_{mo} t \leq 0, \quad m = 1, 2, \dots, M, \\ & \Lambda_{bo} y_{nb} + y_{no} \Psi_n - y_{no} t \geq 0, \quad n = 1, 2, \dots, N, \\ & \Phi_m \geq 0, \quad \Psi_n \geq 0, \quad \Lambda_{bo} \geq 0. \end{aligned} \tag{29}$$

In the second stage, the DMUs are ranked according to the technical efficiency indices (TE). For DMU_o , such an index is defined as an arithmetic mean of the optimal values of θ_o^b obtained from the above model:

$$TE_o = \frac{\sum_{k=1}^K \theta_o^k}{K}. \tag{30}$$

As far as the illustrative example is concerned, the minimum distances of DMU_o to the point in the intersection of the projection set of DMU_o and benchmark DMU_b (θ_o^b) are presented in Table 11. The technical efficiency (TE_o), computed as an average of these distances, and the underlying ranks are given in the last two columns. Such a score is very high for unit D, which attains the first rank. It derives from great efficiencies computed when some inefficient units (e.g., C or E) are set as the benchmarks. On the contrary, the technical efficiency is close to one for unit C, which attains a unitary score for seven out of ten considered scenarios.

3.5. MCDA-based method (AHP-DEA)

The next group of methods for ranking DMUs is inspired by MCDA. We refer to the adaptation of the Analytic Hierarchy Process (AHP) (Abastante et al., 2019; Saaty, 1980), which was proposed by Sinuany-Stern et al. (2000). The idea is to evaluate the cross-efficiency for each pair of DMUs (DMU_a, DMU_b). The cross-efficiency of DMU_a using the weights optimal for DMU_b is marked as E_{ab} . Having found four cross-efficiency scores: $E_{aa}, E_{ab}, E_{ba}, E_{bb}$ for each pair (DMU_a, DMU_b) $\in \mathcal{D} \times \mathcal{D}$, we construct comparison matrix M with the elements defined in the following way:

$$\begin{aligned} a_{aa} &= 1, & a &= 1, 2, \dots, K, \\ a_{ab} &= \frac{E_{aa} + E_{ab}}{E_{bb} + E_{ba}} & a &= 1, 2, \dots, K, b = 1, 2, \dots, K, a \neq b. \end{aligned} \tag{31}$$

In the second step, matrix M is exploited with AHP. Specifically, we find the eigenvector w corresponding to the maximal eigenvalue λ_{max} of matrix M . Such an eigenvector contains the priorities of all DMUs. The DMU corresponding to the greatest value in w is ranked at the top, and other DMUs are ordered according to the descending values of w .

The matrix of cross-efficiency scores derived from the analysis of all pairs of DMUs in the illustrative example is shown in Table 12. For example, cross-efficiency for unit C using weights optimal for unit B is equal to 0.564. In contrast, unit B's cross-efficiency using weights

Table 11

Optimal distances of DMUs to the projection sets of other DMUs (θ_o^b), values of technical efficiency (TE_o) and ranks for all DMUs in the illustrative example.

Unit	A	B	C	D	E	F	G	H	I	J	TE_o	Rank BI
A	1.000	1.853	8.218	2.160	4.289	1.738	1.162	3.774	10.997	2.082	3.727	4
B	2.317	1.000	18.961	1.625	5.873	2.377	1.623	4.297	6.650	2.488	4.721	2
C	1.000	1.000	1.000	1.000	1.086	1.000	1.000	1.815	1.909	1.000	1.181	10
D	3.643	1.749	87.242	1.000	26.146	6.175	2.036	4.424	4.356	6.287	14.306	1
E	1.000	1.000	2.182	1.000	1.000	1.000	1.000	1.859	1.926	1.000	1.297	9
F	1.387	1.454	10.614	1.426	5.012	1.000	1.330	1.958	2.664	1.289	2.813	5
G	1.503	1.995	8.003	1.917	3.644	1.837	1.000	6.838	16.974	1.888	4.560	3
H	1.000	1.007	1.973	1.630	1.935	1.499	1.000	2.024	1.514	1.458	1.458	7
I	1.000	1.000	1.973	1.061	1.630	1.504	1.000	1.030	1.000	1.479	1.325	8
J	1.338	1.239	8.994	1.171	3.788	1.033	1.279	1.938	2.205	1.000	2.399	6

Table 12
Cross efficiency matrix derived from the analysis of all pairs of DMUs for the illustrative example.

Unit	A	B	C	D	E	F	G	H	I	J
A	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
B	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
C	0.246	0.564	1.000	0.344	1.000	0.212	0.342	1.000	1.000	0.346
D	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
E	0.850	0.699	1.000	0.426	1.000	0.375	0.745	1.000	1.000	0.429
F	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
G	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
H	0.879	1.000	1.000	1.000	1.000	1.000	0.326	1.000	1.000	1.000
I	0.889	0.286	1.000	1.000	1.000	1.000	0.330	1.000	1.000	1.000
J	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 13
Comparison matrix for the AHP-DEA ranking approach, values w_i contained in the corresponding principal eigenvector, and final ranks of DMUs in the illustrative example.

Unit	A	B	C	D	E	F	G	H	I	J	w_i	Rank AHP-DEA
A	1.000	1.000	4.058	1.000	1.177	1.000	1.000	1.138	1.125	1.000	0.113	4
B	1.000	1.000	1.774	1.000	1.431	1.000	1.000	1.000	3.499	1.000	0.120	3
C	0.246	0.564	1.000	0.344	1.000	0.212	0.342	1.000	1.000	0.346	0.050	10
D	1.000	1.000	2.910	1.000	2.347	1.000	1.000	1.000	1.000	1.000	0.113	5
E	0.850	0.699	1.000	0.426	1.000	0.375	0.745	1.000	1.000	0.429	0.067	9
F	1.000	1.000	4.721	1.000	2.665	1.000	1.000	1.000	1.000	1.000	0.124	2
G	1.000	1.000	2.926	1.000	1.342	1.000	1.000	3.063	3.027	1.000	0.138	1
H	0.879	1.000	1.000	1.000	1.000	1.000	0.326	1.000	1.000	1.000	0.085	7
I	0.889	0.286	1.000	1.000	1.000	1.000	0.330	1.000	1.000	1.000	0.077	8
J	1.000	1.000	2.890	1.000	2.332	1.000	1.000	1.000	1.000	1.000	0.113	6

optimal for unit C is equal to 1. Then, we construct the comparison matrix (see Table 13). The principal eigenvector derived from its analysis and the corresponding ranks based on the priorities contained in the results obtained with AHP are provided in the last two columns. All efficient units have priorities of at least 0.113. In contrast, the least priority of 0.050 is assigned to unit C. In particular, the highest score of unit G follows the favorable results of its comparison with the inefficient units such as C, H, and I.

3.6. Network-based DEA (NDEA)

A network-based approach for ranking DMUs follows a five-step procedure (Liu and Lu, 2010). The idea is based on investigating the role of each DMU as a benchmark for another DMU, considering all possible input-output combinations. The five steps are as follows:

- Step 1: Each unit is considered as a node in the network. The nodes are connected with the directed links with weights equal to the values of λ_{ko} from the dual formulation of the standard DEA model.
- Step 2: The standard DEA model is applied for multiple problem specifications. A single specification (t) corresponds to one input/output combination. The DEA model is run for all input/output combinations.

Table 14
Adjacency matrix for the network obtained in the illustrative example.

Unit	A	B	C	D	E	F	G	H	I	J
A	0.000	0.822	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.083
B	2.000	0.000	1.000	2.000	1.286	1.243	2.000	2.514	4.000	1.548
C	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
D	1.000	1.592	2.000	0.000	2.375	1.866	1.000	1.147	1.615	2.137
E	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
F	2.000	1.586	4.000	2.000	3.625	0.000	2.000	1.853	1.385	3.460
G	2.000	1.000	2.000	1.000	1.714	1.890	0.000	3.486	2.000	1.773
H	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
I	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
J	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Step 3: The values λ_{ko}^t obtained in the previous step are normalized. It prevents smaller DMUs from receiving greater λ_{ko}^t values than the bigger DMUs. The normalization is performed as follows:

$$IW_{mi}^{t,o} = \frac{\lambda_{io}^t x_{mi}}{\sum_{k=1}^K \lambda_{ko}^t x_{mk}} \quad \text{and} \quad OW_{ni}^{t,o} = \frac{\lambda_{io}^t y_{ni}}{\sum_{k=1}^K \lambda_{ko}^t y_{nk}} \quad (32)$$

A normalized overall share of contribution of DMU_i in the reference set of DMU_o is computed using the following formula:

$$IOW_{oi}^t = \frac{\sum_{m=1}^M IW_{mi}^{t,o} + \sum_{n=1}^N OW_{ni}^{t,o}}{N + M} \quad (33)$$

Step 4: The network is constructed by aggregating the results for all w specifications into a single network represented as an adjacency matrix A with elements defined as:

$$a_{ij} = \left[\sum_{t=1}^w IOW_{ij}^t \right] \quad (34)$$

Step 5: The eigenvector v corresponding to the greatest eigenvalue of A is computed, and the DMUs are ranked according to the descending values of v .

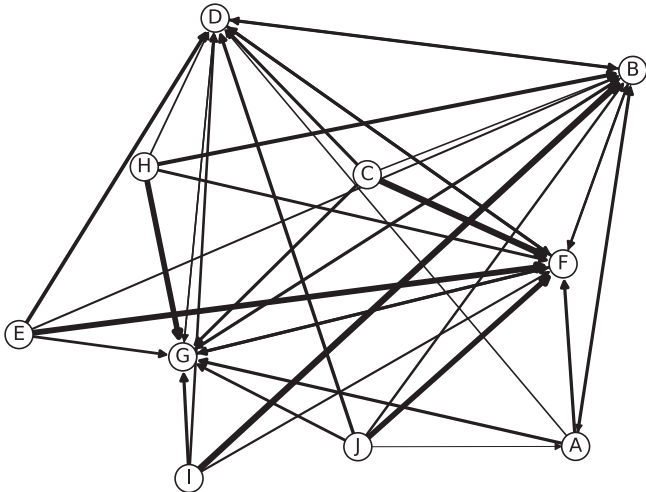


Fig. 1. The network obtained with the network-based DEA method for the illustrative example (thicker lines correspond to greater weights).

Table 15

Priorities w_i of the efficient DMUs derived from the principal eigenvector of the adjacency matrix and the final ranks obtained in the network-based DEA for the illustrative example.

Unit	w_i	NDEA Rank
A	0.085	5
B	0.525	2
D	0.470	3
F	0.554	1
G	0.435	4

The adjacency matrix for the network constructed after the analysis of all different specifications, normalization, and aggregation is presented in Table 14. The corresponding graph is given in Fig. 1. The principal eigenvector of the adjacency matrix and the ranking of efficient DMUs are provided in Table 15. The priorities assigned to B, D, F, and G are 4–5 times greater than A’s priority.

3.7. Methods based on robustness analysis

The methods based on RA exploit the results derived with the Monte Carlo simulation (Kadziński et al., 2017; Ciomek & Kadziński, 2021). They take advantage of the multiplicity of input/output weights that can serve as the basis for comparing DMUs. For each weight vector (u^i, v^j) from a large representative subset S of input/output weights sampled from a uniform distribution, we compute the efficiency score E_o^i for DMU_o :

$$E_o^i = \frac{\sum_{n=1}^N u_n^i y_{no}}{\sum_{m=1}^M v_m^j x_{mo}} \tag{35}$$

The results obtained for all analyzed samples are summarized or exploited in four different ways. Two of them have been originally elaborated by Lahdelma and Salminen (2006) and Kadziński et al. (2017), whereas the other two approaches are introduced in the context of DEA in this paper. Overall, these ranking methods exploit the results which are derived from one vs. one or one against all comparisons, building on the three relevant perspectives on the robustness of DMUs’ efficiency, i.e., scores, attained ranks, and pairwise preference relations.

First, as originally proposed by Lahdelma and Salminen (2006), we may compute the expected efficiency (EE) over all weight vectors:

$$EE_o = \frac{\sum_{i=1}^{|S|} E_o^i}{|S|} \tag{36}$$

Second, instead of analyzing the efficiency scores, we may focus on the ranks. For DMU_o efficiency rank acceptability index $ERAI(DMU_o, k)$ quantifies the share of weights for which DMU_o is ranked k -th. Then, an expected efficiency rank (ER) can be computed as follows (Kadziński et al., 2017):

$$ER_o = \sum_{k=1}^K k \cdot ERAI(DMU_o, k) \tag{37}$$

The remaining two methods exploit the pairwise efficiency out-ranking indices $PEOI(DMU_o, DMU_k)$ that capture the share of weights for which DMU_o attains at least as good efficiency as DMU_k . The matrix of $PEOIs$ may be exploited in a twofold way, inspired by different MCDA methods. For this reason, they may also be classified with the category of MCDA-inspired approaches. On the one hand, we may adapt the idea initially implemented within the Net Flow Score (NFS) procedures used, e.g., in the PROMETHEE methods (Kadziński and Michalski, 2016). Let us denote this method by NFS-PEOI. Specifically, for each DMU, we may derive its net flow $\Phi(DMU_o)$ as a difference between its positive Φ^+ and negative Φ^- flows. The positive flow quantifies the comprehensive strength of DMU_o as its advantage over all remaining DMUs in terms of $PEOIs$. In contrast, the negative flow captures the comprehensive weakness of DMU_o compared with all other units. In this perspective, an overall measure of desirability NFS_o quantifies the difference between shares of feasible input/output weights for which DMU_o is ranked at least as good and not better as other units:

$$NFS_o = \sum_{k=1}^K [PEOI(DMU_o, DMU_k) - PEOI(DMU_k, DMU_o)] \tag{38}$$

On the other hand, within the PEV-PEOI (Principal Eigenvector-based exploitation of $PEOIs$) method, the matrix of $PEOIs$ is exploited using the eigenvector method (Saaty, 1980). Even if the exploitation procedure is the same as in AHP-DEA and NDEA, the results’ interpretation is very different. The derived priorities – corresponding to the values in the principal eigenvector of the $PEOI$ matrix – capture each DMU’s importance given the shares of feasible weights for which it is better and worse than other units. Hence, they are strictly linked to the robustness concern. The main difference between PEV-PEOI and NFS-PEOI comes from the fact that the former derives a score for each DMU while already accounting for the quality of other DMUs. In contrast, the latter sums up the arguments in favor and against a given DMU irrespective of the status of other DMUs that proved to be better or worse from the considered unit.

To illustrate the methods based on Robustness Analysis, we analyze only five vectors of input/output weights (see Table 16). In practice, the number of such vectors is usually a few thousand. For each weight vector, we compute the efficiency scores. As they are not necessarily laying in the interval $[0, 1]$, we normalize them through dividing by the maximal value obtained for a given weight vector. In Table 17, we present the efficiencies and ranks for each weight vector. Such elementary results are averaged over all samples to derive the expected efficiencies and ranks. For example, when analyzing only the outcomes for the five generated sample, unit I attains an average efficiency of 0.058, hence being ranked at the very bottom according to EE , and an average rank of 9, which places it in the ninth position in terms of ER .

The matrix of $PEOIs$ is presented in Table 18. For example, unit A is ranked at least as good as unit B for all studied weight vectors ($PEOI(A, B) = 1$). In case of units B and F, the former attains a higher efficiency

Table 16

A sample of five input/output weights derived with the Monte Carlo simulation.

Sample	v_1	v_2	u_1	u_2
1	0.900	0.100	0.252	0.748
2	0.769	0.231	0.448	0.552
3	0.974	0.026	0.754	0.246
4	0.942	0.058	0.356	0.644
5	0.600	0.400	0.631	0.369

Table 17

The efficiency scores and ranks obtained for five generated samples, expected efficiency scores (EE), expected ranks (ER) and ranks obtained by the DMUs according to these two measures (Rank EE) and (Rank ER) for the illustrative example.

Unit	Sample					EE	Rank EE	ER	Rank ER
	1	2	3	4	5				
A	0.796 (2)	0.618 (2)	0.617 (2)	0.744 (2)	0.512 (2)	0.657	2	2.0	2
B	0.386 (6)	0.368 (4)	0.229 (3)	0.321 (4)	0.408 (3)	0.342	4	4.0	4
C	0.11 (9)	0.054 (10)	0.026 (10)	0.083 (9)	0.028 (10)	0.060	9	9.6	10
D	0.433 (4)	0.309 (5)	0.101 (7)	0.301 (5)	0.262 (4)	0.281	5	5.0	5
E	0.161 (7)	0.094 (8)	0.03 (9)	0.112 (8)	0.059 (9)	0.091	8	8.2	8
F	0.631 (3)	0.384 (3)	0.173 (5)	0.466 (3)	0.256 (5)	0.382	3	3.8	3
G	1.000 (1)	1.000 (1)	1.000 (1)	1.000 (1)	1.000 (1)	1.000	1	1.0	1
H	0.153 (8)	0.195 (7)	0.196 (4)	0.167 (7)	0.231 (6)	0.188	7	6.4	7
I	0.045 (10)	0.062 (9)	0.052 (8)	0.047 (10)	0.085 (8)	0.058	10	9.0	9
J	0.409 (5)	0.267 (6)	0.117 (6)	0.3 (6)	0.195 (7)	0.258	6	6.0	6

Table 18

Pairwise efficiency outranking indices (PEOIs) for all pairs of DMUs in the considered example.

Unit	A	B	C	D	E	F	G	H	I	J
A	1.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0	1.0	1.0
B	0.0	1.0	1.0	0.8	1.0	0.4	0.0	1.0	1.0	0.8
C	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0
D	0.0	0.2	1.0	1.0	1.0	0.2	0.0	0.8	1.0	0.8
E	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.2	0.6	0.0
F	0.0	0.6	1.0	0.8	1.0	1.0	0.0	0.8	1.0	1.0
G	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
H	0.0	0.0	1.0	0.2	0.8	0.2	0.0	1.0	1.0	0.4
I	0.0	0.0	0.6	0.0	0.4	0.0	0.0	0.0	1.0	0.0
J	0.0	0.2	1.0	0.2	1.0	0.0	0.0	0.6	1.0	1.0

Table 19

Scores and ranks derived with PEV-PEOI and NFS-PEOI for the illustrative example based on the matrix of PEOIs.

Unit	PEV-PEOI Score	Rank PEV-PEOI	NFS-PEOI Score	Rank NFS-PEOI
A	0.4866	2	7.0	2
B	0.2276	4	3.0	4
C	0.0007	10	-8.2	10
D	0.1447	5	1.0	5
E	0.0108	8	-5.4	8
F	0.2506	3	3.4	3
G	0.7842	1	9.0	1
H	0.0760	7	-1.8	7
I	0.0029	9	-7.0	9
J	0.0822	6	-1.0	6

for 40% of feasible weights ($PEOI(B, G) = 0.4$), whereas an inverse relation holds for 60% of input/output weights ($PEOI(F, B) = 0.6$).

We exploit the matrix of PEOIs in two different ways. First, we compute the Net Flow Scores for each DMU. In this regard, a detailed analysis for unit B can be conducted as follows:

$$NFS_B = (0 + 1 + 1 + 0.8 + 1 + 0.4 + 0 + 1 + 1 + 0.8) - (1 + 1 + 0 + 0.2 + 0 + 0.6 + 1 + 0 + 0 + 0.2) = 7 - 4 = 3. \tag{39}$$

It sums up the row's values corresponding to unit B in the PEOI matrix and subtracts the values from the respective column. A positive score for B means that the share of feasible weight vectors for which it attained a better score than other DMUs is greater than the share of input/output weights for which it proved to be worse. The detailed scores and ranks derived with this method are presented in Table 19. For example, unit G attains the highest possible score because it turned out to be better than all remaining DMUs for all weight vectors. On the other extreme, unit C's score is very negative because the results of its comparison with all remaining units are unfavorable.

For the scores and ranks derived with PEV-PEOI, see Table 19. For this particular problem, the ranks are the same as those obtained with

NFS-PEOI. However, the scores differ a lot. First, they are all positive, whereas in NFS-PEOI, the least advantageous units attain the negative scores. Second, their interpretation is relative rather than absolute. That is, in NFS-PEOI, the score can be interpreted on a scale between the greatest and the least possible values, whereas in PEV-PEOI, one should rather focus on the ratios of scores attained by different pairs of units. Third, PEV-PEOI tends to emphasize the evident advantage of some units over the others observed in the matrix of PEOIs. For example, when comparing the top-ranked units G and A, the score assigned to G is by over 60% higher. This is mainly due to its superiority over unit A demonstrated by $PEOI(G, A) = 1$. In NFS-PEOI, the advantage of G over A also derives from the direct comparison of these two units. However, it is smaller in magnitude because both these units compare positively with all remaining ones.

3.8. Virtual DMU method (VDMU)

The last presented method (Wang and Luo, 2006) is inspired by another MCDA approach, called TOPSIS (Hwang and Yoon, 1981; de Lima Silva et al., 2020). It is representative of the group of procedures incorporating single or multiple artificial, invisible, dummy, or virtual DMUs in the ranking procedure (see, e.g., Hosseinzadeh Lotfi et al., 2011; Jahanshahloo et al., 2010; Kritikos, 2017; Wang and Yang, 2007). Specifically, it accounts for a pair of virtual DMUs, ideal (IDMU) and anti-ideal (ADMU) ones. IDMU consumes the least possible inputs and produces the greatest outputs:

$$\begin{aligned} x_{mi} &= \min_k \{x_{mk}\} & m = 1, 2, \dots, M, k = 1, 2, \dots, K, \\ y_{ni} &= \max_k \{y_{nk}\} & n = 1, 2, \dots, N, k = 1, 2, \dots, K. \end{aligned} \tag{40}$$

In turn, ADMU consumes the greatest inputs and produces the least outputs:

$$\begin{aligned} x_{mA} &= \max_k \{x_{mk}\} & m = 1, 2, \dots, M, k = 1, 2, \dots, K, \\ y_{nA} &= \min_k \{y_{nk}\} & n = 1, 2, \dots, N, k = 1, 2, \dots, K. \end{aligned} \tag{41}$$

The super-efficiency model is applied for the basic dataset enriched with IDMU. Specifically, the efficiency of an ideal unit is maximized:

$$\begin{aligned} \max \quad & \theta_{IDMU} = \sum_{n=1}^N u_n y_{ni} \\ \text{subject to:} \quad & \sum_{m=1}^M v_m x_{mi} = 1, \\ & \sum_{n=1}^N u_n y_{nk} \leq \sum_{m=1}^M v_m x_{mk}, \quad k = 1, 2, \dots, K, \\ & u_n \geq 0, v_m \geq 0, \quad n = 1, 2, \dots, N; m = 1, 2, \dots, M. \end{aligned} \tag{42}$$

Since the above model can have multiple optimal solutions, the

following model is subsequently applied for each $DMU_o \in \mathcal{D}$:

$$\begin{aligned}
 \max \quad & \theta_o = \sum_{n=1}^N u_n y_{no} \\
 \text{subject to:} \quad & \sum_{m=1}^M v_m x_{mo} = 1, \\
 & \sum_{n=1}^N u_n y_{nk} \leq \sum_{m=1}^M v_m x_{mk}, \quad k = 1, 2, \dots, K, \\
 & \sum_{n=1}^N u_n y_{nl} = \theta_{IDMU}^* \sum_{m=1}^M v_m x_{ml}, \\
 & u_n \geq 0, v_m \geq 0, \quad n = 1, 2, \dots, N; m = 1, 2, \dots, M,
 \end{aligned} \tag{43}$$

where θ_{IDMU}^* is the optimal objective value from model (42). Hence, $IDMU$ is constrained to attain the greatest possible efficiency θ_{IDMU}^* . The optimal value for the model above is denoted by θ_o^* . A similar analysis is conducted for the dataset enriched with $ADMU$, whose efficiency is minimized. Specifically, we consider the following two models:

$$\begin{aligned}
 \min \quad & \phi_{ADMU} = \sum_{n=1}^N u_n y_{nA} \\
 \text{subject to:} \quad & \sum_{m=1}^M v_m x_{mA} = 1, \\
 & \sum_{n=1}^N u_n y_{nk} \geq \sum_{m=1}^M v_m x_{mk}, \quad k = 1, 2, \dots, K, \\
 & u_n \geq 0, v_m \geq 0, \quad n = 1, 2, \dots, N; m = 1, 2, \dots, M.
 \end{aligned} \tag{44}$$

$$\begin{aligned}
 \min \quad & \phi_o = \sum_{n=1}^N u_n y_{no} \\
 \text{subject to:} \quad & \sum_{m=1}^M v_m x_{mo} = 1, \\
 & \sum_{n=1}^N u_n y_{nk} \geq \sum_{m=1}^M v_m x_{mk}, \quad k = 1, 2, \dots, K, \\
 & \sum_{n=1}^N u_n y_{nA} = \phi_{ADMU}^* \sum_{m=1}^M v_m x_{mA}, \\
 & u_n \geq 0, v_m \geq 0, \quad n = 1, 2, \dots, N; m = 1, 2, \dots, M.
 \end{aligned} \tag{45}$$

The optimal solutions of models (44) and (45) are denoted by, respectively, ϕ_{ADMU}^* and ϕ_o^* . The DMUs are ranked according to the relative closeness (RC) index defined as follows:

$$RC_o = \frac{\phi_o^* - \phi_{ADMU}^*}{(\phi_o^* - \phi_{ADMU}^*) + (\theta_{IDMU}^* - \theta_o^*)}. \tag{46}$$

The inputs and outputs for the ideal and anti-ideal virtual units for the illustrative example are presented in Table 20. The maximal efficiency score for the ideal unit $IDMU$ is $\theta_{IDMU}^* = 13.843$, and the minimal efficiency for the anti-ideal unit is $\phi_{ADMU}^* = 0.09$.

For each DMU_o , the efficiency scores relative to the ideal and anti-ideal DMUs (ϕ_o^* and θ_o^*) are presented in Table 21. Finally, the units are ordered according to the relative closeness scores (RC_o). The highest value is attained by unit G, whereas the lowest one by unit I. Interestingly, efficient unit D is ranked lower than two inefficient units H and J.

3.9. Features of the considered ranking methods

In Table 22, we summarize the strengths and weaknesses of the presented ranking methods. They can serve as a guide supporting

Table 20
Input and output values for virtual ideal ($IDMU$) and anti-ideal ($ADMU$) DMUs.

Virtual Unit	i_1	i_2	o_1	o_2
$IDMU$	73	53	852	919
$ADMU$	910	983	12	11

Table 21

Efficiency scores in relation to the ideal and anti-ideal units, relative closeness scores and final ranks of DMUs according to the virtual DMU method for the illustrative example.

Unit	ϕ_o^*	θ_o^*	RC_o	Rank VDMU
A	0.805	14.333	0.522	2
B	0.677	5.192	0.279	4
C	0.138	1.000	0.062	9
D	1.000	3.324	0.201	7
E	0.290	1.180	0.074	8
F	1.000	5.656	0.302	3
G	1.000	21.004	0.619	1
H	0.135	3.823	0.214	5
I	0.053	1.000	0.061	10
J	0.709	3.658	0.213	6

selecting a ranking method based on its features. In particular, we refer to the number of weight vectors considered when constructing a ranking, a subset of units affected by ranking with a given method, computational complexity and techniques, intuitiveness of the provided scores, and some peculiar characteristics that cannot be generalized to all methods.

3.10. Summary of the rankings derived with different methods for the illustrative example

In Table 23, we present the rankings obtained with sixteen considered methods (including the standard CCR model). Let us emphasize that the inefficient units are ordered according to their CCR efficiency scores for the approaches that discriminate only between the efficient ones. When it comes to the very top of rankings, the first position is attained by G, D, and A for, respectively, 11, 4, and 2 methods. In turn, C and I are ranked at the very bottom by, respectively, 13 and 3 procedures. The positions attained by the inefficient units for various methods are relatively stable. For example, J is always ranked sixth, E attains positions between 7 and 9, whereas C is ranked at the two bottom ranks by all procedures. When it comes to efficient units, the variety of attained ranks is greater. In particular, A is ranked first in the best possible scenarios and fifth in the worst case, whereas the ranks of G are in the interval [1, 4]. D is the only efficient unit ranked worse than some inefficient units for some ranking methods (see VDMU, where H and J are ranked better than E).

4. Experimental comparison of rankings provided by different methods

This section aims to demonstrate how the same problem (i.e., a ranking of DMUs with the same input and output values) can be approached by methods representing different streams in DEA. Specifically, we report the results of an extensive experimental comparison of ranking methods in DEA in terms of the results they provide. For this purpose, we consider artificial and real-world datasets, as well as five similarity measures. However, we do not indicate a clear winner nor good or bad methods. We neither claim that there are clear benefits in favor of one approach compared to the others. Unlike in machine learning, there is no objective truth to be attained in DEA, and the assumption that the DM's true ranking pre-exists is unrealistic. Hence, the experimental comparison of different ranking methods with objective reality is ill-founded. Nevertheless, the empirical comparison of the outputs of different DEA ranking methods is meaningful because there exists a common context of their use. Numerous ranking methods have been proposed over the last decades. Each of them is based on different axioms and introduces some instrumental bias in its steps, potentially leading to different results. Our experiments constitute an important step in verifying the similarity of outputs provided by several ranking methods. We want to reveal analogies and differences between these

Table 22
Main advantages and disadvantages of the considered ranking methods.

Method	Advantages	Disadvantages
CE	Multiple weight vectors considered. Drops unrealistic weight schemes. Applies peer and unbiased self-evaluation.	Limited set of common weights. Ambiguity in the selection of weights.
SE	Simplicity. Detecting outliers.	Ranks only efficient units. Occasional infeasibility. No common basis for the comparison of units. Can favor specialized units.
CCA	Common basis for the comparison of units. Ranks all units.	Reliance on a single weight vector. Inefficient unit can be ranked at the top. Occasional infeasibility. Complex application. Sensitivity of eigenvector computation.
LDA	Common basis for the comparison of units. Ranks all units.	Reliance on a single weight vector. Inefficient unit can be ranked at the top. Occasional infeasibility.
DR-DEA	Common basis for the comparison of units. Ranks all units.	Reliance on a single weight vector. Inefficient unit can be ranked at the top.
BSA	Investigates the impact of efficient units on the inefficient ones.	Ranks only efficient units. Complex interpretation of scores.
BCRS	Investigates the impact of efficient units on the inefficient ones. Multiple weight vectors considered. Simple and direct application.	Ranks only efficient units. Limited set of common weights considered.
BI	Investigates the impact of units on the efficiency of others. Ranks all units.	No common basis for the comparison of units.
AHP-DEA	Incorporates DMUs' cross-efficiency comparisons. Ranks all units.	Inefficient unit can be ranked at the top. Sensitivity of eigenvector computation.
NDEA	Considers multiple input-output settings. Ranks all units.	High time complexity. Sensitivity of eigenvector computation. Complex interpretation of scores.
EE	All feasible weight vectors considered. Avoids arbitrary selection of weights. Intuitive interpretation of scores. Ranks all units.	Requires sampling procedure. Inefficient unit can be ranked at the top.
ER	All feasible weight vectors considered. Avoids arbitrary selection of weights. Intuitive interpretation of scores. Ranks all units.	Requires sampling procedure. Averages ordinal measures (ranks). Inefficient unit can be ranked at the top.
PEV-PEOI	All feasible weight vectors considered. Avoids arbitrary selection of weights. Based on DMUs' pairwise comparisons. Ranks all units.	High time complexity. Requires sampling procedure. Sensitivity of eigenvector computation. Inefficient unit can be ranked at the top.
NFS-PEOI	All feasible weight vectors considered. Avoids arbitrary selection of weights. Based on DMUs' pairwise comparisons. Ranks all units.	Requires sampling procedure. Inefficient unit can be ranked at the top.
VDMU		Changes the original set of DMUs. High sensitivity to outlying DMUs

Table 22 (continued)

Method	Advantages	Disadvantages
	Simplicity and intuitiveness. Low time complexity. Ranks all units.	due to incorporating extreme units. Limited set of weight vectors. Ambiguity in the selection of weights.

Table 23

Ranks attained by the ten DMUs according to sixteen ranking methods applicable in the context of DEA for the illustrative example.

Method	A	B	C	D	E	F	G	H	I	J
CCR	1	1	10	1	8	1	1	7	9	6
CE	5	3	10	1	7	4	2	8	9	6
SE	5	4	10	1	8	3	2	7	9	6
CCA	3	4	10	2	8	5	1	7	9	6
LDA	1	4	10	3	7	5	2	8	9	6
DR-DEA	5	4	9	2	7	3	1	8	10	6
BSA	5	2	10	4	8	3	1	7	9	6
BCRS	5	4	10	3	8	2	1	7	9	6
BI	4	2	10	1	9	5	3	7	8	6
AHP-DEA	4	3	10	5	9	2	1	7	8	6
NDEA	5	2	10	3	8	1	4	7	9	6
EE	2	4	9	5	8	3	1	7	10	6
ER	2	4	10	5	8	3	1	7	9	6
PEV-PEOI	2	4	10	5	8	3	1	7	9	6
NFS-PEOI	2	4	10	5	8	3	1	7	9	6
VDMU	2	4	9	7	8	3	1	5	10	6

approaches while referring to the results they provide. Such conclusions can be used in a twofold way. On the one hand, in real-world decision analysis, one can select a single approach from a sub-group providing the same or very similar results on a broad spectrum of problems. Even if the underlying rules of these methods differ, their recommendations match to a large extent. In this way, the user is guided in the plethora of available methods. On the other hand, to investigate different aspects of the problem, one can apply approaches from sub-groups whose application on the same problem instances leads to dissimilar results. This is useful for understanding the conditions under which different units can be judged more or less preferred than others. Obviously, these findings need to be confronted against potential users' subjective opinions because the usefulness of adopting some ideas in practice is the matter of the reader's individual interpretation that should also consider the strength and weaknesses discussed in Section 3.9.

4.1. Choice and ranking correlation measures

To compare the rankings generated by different methods, we considered the following choice and ranking similarity measures (Kadziński and Michalski, 2016):

- **Hit ratio (HR)** (Barron and Barrett, 1996) – a binary measure which is equal to one if both methods rank the same DMU at the top, which can be represented with the following formula:

$$HR(R_1, R_2) = \begin{cases} 1 & \text{if } R_1(1) \cap R_2(1) \neq \emptyset \\ 0 & \text{otherwise,} \end{cases} \quad (47)$$

where $R_x(r)$ is a set of DMUs ranked r -th in the order imposed by the R_x method (e.g., $R_1(1)$ is a set of DMUs ranked at the very top by procedure R_1). For example, based on the results presented in Section 3.10, $HR(CE, SE) = 1$ as both CE and SE rank D at the top, whereas $HR(CCA, LDA) = 0$, because CCA and LDA rank, respectively, G and A at the first position. The examples for the remaining measures are also based on the outcomes provided in Section 3.10.

- **Normalized Hit Ratio (NHR)** (Kadziński and Michalski, 2016) is an extension of the HR measure, which takes into consideration the partial agreement between rankings. It is defined as follows:

$$NHR(R_1, R_2) = \frac{R_1(1) \cap R_2(1)}{R_1(1) \cup R_2(1)}. \tag{48}$$

When both methods rank the same set of DMUs at the top, then $NHR(R_1, R_2) = HR(R_1, R_2) = 1$ (e.g., $NHR(CE, SE) = 1$). If some DMU is ranked the best using both procedures, but the subsets of top units are not the same ($R_1(1) \neq R_2(1)$), then $HR(R_1, R_2) = 1$ and $NHR(R_1, R_2) < 1$ (e.g., $NHR(CCR, CE) = 0.2$ as CCR ranks five units at the top, and one of them is D, which is also judged as the most preferred by CE).

- **Kendall's τ** (Hays and Winkler, 1970) is derived from the analysis of agreements and disagreements between relations observed in the two rankings for all pairs of DMUs. Let us denote the preference and indifference relations in the ranking provided by R by, respectively, \succ^R and \sim^R . A function $p(R, DMU_o, DMU_k)$ translating the relation observed for a pair (DMU_o, DMU_k) of DMUs into a numerical value is defined as follows:

$$p\left(R, DMU_o, DMU_k\right) = \begin{cases} 1, & \text{if } DMU_o \succ^R DMU_k, \\ 0.5, & \text{if } DMU_o \sim^R DMU_k, \\ 0, & \text{if } DMU_o = DMU_k \vee DMU_k \succ^R DMU_o. \end{cases} \tag{49}$$

Then, Kendall's τ is defined in the following way:

$$\tau\left(R_1, R_2, K\right) = 1 - 4 \frac{d_k(R_1, R_2)}{K \cdot (K - 1)}, \tag{50}$$

where $d_k(R_1, R_2)$ is a Kendall's distance between R_1 and R_2 :

$$d_k\left(R_1, R_2\right) = 0.5 \sum_{(DMU_o, DMU_k) \in \mathcal{D} \times \mathcal{D}} \left| p\left(R_1, DMU_o, DMU_k\right) - p\left(R_2, DMU_o, DMU_k\right) \right|. \tag{51}$$

The extreme values of $\tau(R_1, R_2, K)$ are 1 and -1 , indicating that the two rankings are the same or one ranking negates all pairwise relations observed in the other one. For example, $\tau(BI, AHP-DEA, 10) = 0.778$, hence being relatively high, because the relations observed in BI and AHP-DEA are different only for 10 out of 90 pairs of DMUs.

- **Rank Difference Measure (RDM)** (Kadziński and Michalski, 2016) takes into account the difference between positions attained by each DMU in the generated rankings. To account for the shared ranks, we consider the best $r^*(R, DMU_o)$ and the worst $r_*(R, DMU_o)$ ranks of DMU_o in ranking R . They can be computed in the following way:

$$r^*(R, DMU_o) = K - |DMU_k, k = 1, \dots, K, k \neq o : DMU_o \succ^R DMU_k|, \tag{52}$$

$$r_*(R, DMU_o) = 1 + |DMU_k, k = 1, \dots, K, k \neq o : DMU_k \succ^R DMU_o|, \tag{53}$$

where $\succ^R = \succ^R \cup \sim^R$ is a weak preference relation interpreted as "being at least as good" in ranking R . For example, when considering the illustrative example, the best rank r^* of units A, B, D, F, and G in the ranking obtained by applying the CCR method is one because they are at least as good as nine other DMUs ($10 - 9 = 1$). On the contrary, their worst rank r_* is the fifth position ($1 + 4 = 5$) because four other units are ranked at least as good. Then, the RDM measure is defined as follows:

$$RDM\left(R_1, R_2, K\right) = 1 - \frac{\sum_{DMU_o \in \mathcal{D}} |\bar{r}(R_1, DMU_o) - \bar{r}(R_2, DMU_o)|}{\max_{\text{diff}}^{\text{rank}}(K)}, \tag{54}$$

where $|\bar{r}(R_1, DMU_o) - \bar{r}(R_2, DMU_o)|$ is an average distance between positions attained by DMU_o in the two rankings, R_1 and R_2 :

$$\begin{aligned} & \left| \bar{r}\left(R_1, DMU_o\right) - \bar{r}\left(R_2, DMU_o\right) \right| \\ &= \frac{\sum_{r_1=r^*(R_1, DMU_o)}^{r_*(R_1, DMU_o)} \sum_{r_2=r^*(R_2, DMU_o)}^{r_*(R_2, DMU_o)} |r_1 - r_2|}{(r_*(R_1, DMU_o) - r^*(R_1, DMU_o) + 1) \cdot (r_*(R_2, DMU_o) - r^*(R_2, DMU_o) + 1)} \end{aligned} \tag{55}$$

and

$$\max_{\text{diff}}^{\text{rank}}(K) = \begin{cases} \left\lfloor \frac{K}{2} \right\rfloor \cdot K, & \text{if } K \text{ is even,} \\ \left\lfloor \frac{K}{2} \right\rfloor \cdot (K - 1), & \text{if } K \text{ is odd.} \end{cases} \tag{56}$$

The RDM measure takes values between 0 and 1, where 1 means that the analyzed rankings are the same, and 0 denotes that they are inverse and hence a difference between the positions attained by the DMUs is the highest possible. For example, $RDM(CCA, LDA, 10) = 1 - 6/50 = 0.88$, hence indicating a very high agreement, because the sum of rank differences for all DMUs is equal to six (for a single

DMU, the difference amounts to two, for four DMUs – it is equal to one, and for the remaining five DMUs – it is zero), and in the worst possible case, when considering ten DMUs, the maximal difference between all ranks could be fifty. Note that when comparing the ranking derived with a given procedure with itself, RDM can be lesser than one. This is desired for scenarios when multiple alternatives are ranked ex-aequo, i.e., attain the same rank. In case there are no shared ranks, $RDM(R, R, K) = 1$ for all procedures R and values of K .

- **Rank Acceptance Measure (RAM)** (Kadziński and Michalski, 2016) quantifies how often the same alternative attains the same ranks in both analyzed rankings, hence generalizing NHR to the entire ranking. It is defined as follows:

$$RAM\left(R_1, R_2, K\right) = \frac{1}{K} \sum_{r=1}^K RA\left(R_1, R_2, r\right), \tag{57}$$

where:

$$RA\left(R_1, R_2, r\right) = \frac{|R_1(r) \cap R_2(r)|}{|R_1(r) \cup R_2(r)|}. \tag{58}$$

RAM takes values in the interval $[0, 1]$, where 1 means that all DMUs attain the same position in both rankings, whereas 0 indicates that each DMU attains a different position in the compared rankings. For example, $RAM(ER, VDMU, 10) = 0.6$, because six out of ten DMUs attain the same position in the rankings determined by ER and VDMU.

4.2. Experimental comparison involving artificial datasets

To compare the results of ranking methods discussed in Section 3, we generated 960 random datasets for 96 different problem settings. The datasets involved from 2 to 5 inputs and outputs. To select the number of units, we have studied the typical problem sizes considered in the DEA applications. In particular, we have reviewed a few papers reporting at least several studies and computed some statistics. They indicate that the median number of units in these studies ranges from 23 to 66 (see, e.g., applications on seaports (Panayides et al., 2009) – median of 23, airports (Fasone and Zapata-Aguirre, 2016) – 29, power engineering (Meng et al., 2016) – 30, dynamic DEA (Mariz et al., 2018) – 31.5, healthcare (Kohl et al., 2019) – 53, banking and finances (Kaffash and Marra, 2017) – 57, and education (Worthington, 2001) – 66). The minimal number of units is usually at most ten, whereas the maximum is often a few hundred. Note, however, that in the case of large datasets, it is very rare that one analyzes the whole ranking. In fact, the rank of at most several dozen units is of interest to the user, while the rest of the units are neglected and can remain unordered. For this reason, we have decided to distinguish two size ranges: the most typical, small-scale DEA applications with K ranging from 5 to 30 and large problem instances for which still the entire ranking may be of interest to the user, where K ranges from 75 to 100. Finally, the values of consumed inputs and produced outputs were generated from the following two distributions within the range $[0,1]$: (a) the uniform distribution, and (b) the truncated normal distribution with $\mu = 0.5$ and $\sigma = 0.1$.

For all pairs of methods, we computed five similarity measures. To investigate the potential impact of the number of units K and the distribution of performances, we averaged the results separately for the four scenarios distinguished by the small or large data sets and the uniform or normal performance distribution. In the main paper, we discuss the results for small problem instances with K ranging from 5 to 30 and uniformly distributed performances. The respective results are provided in Tables 24–28. The outcomes for the remaining scenarios are discussed in the e-Appendix (supplementary material available online).

4.2.1. Measures for choice for small problem instances and uniform performance distribution

In Tables 24 and 25, we provide the average results for Hit Ratio and Normalized Hit Ratio. When it comes to HR, let us first focus on the comparisons of CCR with all remaining methods. Its value is equal to one for 11 out of 15 ranking procedures, which means that in all considered problem settings, they rank at the top some unit, which is deemed efficient and hence ranked first with the standard CCR model. The four

methods which rank some inefficient DMU first for some considered settings are CCA (agreement for 63.3% datasets), LDA (99.2%), AHP-DEA (97.9%), and PEV-PEOI (57.9%). The highest agreement in terms of HR is attained by the methods based on Robustness Analysis that incorporate the Monte Carlo simulations to compute the expected ranks, expected efficiencies or net flow scores ($HR(ER, NFS - PEOI) = 0.975$, $HR(EE, NFS - PEOI) = 0.871$, and $HR(ER, EE) = 0.850$) and the two benchmarking methods ($HR(BSA, BCRS) = 0.767$). In general, AHP-DEA attains the highest average similarity to all other methods. For 7 out of 15 methods, its agreement in terms of HR is greater than 0.7. The other three methods which attain high average similarity to all other methods are NFS-PEOI, EE, and ER. On the contrary, CCA attains the least similarity to all other methods, which means that the rules it employs to select the most preferred DMU are very different from these incorporated by the remaining procedures. Three other approaches that score relatively low in terms of HR (i.e., < 0.5) compared with the remaining methods are SE, BI, and PEV-PEOI.

Note that the high similarity of CCR with the remaining ranking procedures in terms of HR results from identifying a subset of DMUs by the CCR model as efficient and equally desirable. This is penalized by NHR, which computes the share of DMUs that are jointly ranked at the top by a pair of methods instead of capturing only a binary agreement. Consequently, the agreement of CCR with all remaining methods in terms of ranking the same DMU at the top quantified with NHR is very low, i.e., ≤ 0.333 . This confirms that ranking methods can discriminate between the efficient units, making them more comparable using some arbitrary measures. In general, NHR values are lower than HR, and when the methods do not rank multiple DMUs at the very top, they are the same or only slightly lesser. The highest average decrease in terms of NHR when compared with HR can be observed for LDA, DR-DEA, and AHP-DEA. The most and the least similar pairs of approaches remained the same as for HR. In particular, CCA is very dissimilar from the remaining procedures in indicating the top-ranked DMU. In contrast, the simulation-based methods NFS-PEOI, EE, and ER capture well, on average, the indications of the most preferred DMU provided by all remaining approaches.

4.2.2. Measures for ranking for small problem instances and uniform performance distribution

In Tables 26–28, we provide the average results for Kendall’s τ , RDM, and RAM. When considering these three similarity measures that take into account the entire rankings, the conclusions in terms of the most similar and dissimilar rankings are consistent to a great extent. Hence, we discuss the conclusions derived from the analysis of Kendall’s τ and

Table 24
Average Hit Ratio for all pairs of ranking procedures over all artificial datasets with a small number of units K and uniform performance distribution.

	CCR	CE	SE	CCA	LDA	DR-DEA	BSA	BCRS	BI	AHP-DEA	NDEA	EE	ER	PEV-PEOI	NFS-PEOI	VDMU
CCR	1.000	1.000	1.000	0.633	0.992	1.000	1.000	1.000	1.000	0.979	1.000	1.000	1.000	0.579	1.000	1.000
CE	1.000	1.000	0.321	0.171	0.508	0.513	0.554	0.554	0.312	0.717	0.425	0.637	0.654	0.308	0.654	0.367
SE	1.000	0.321	1.000	0.150	0.354	0.267	0.246	0.279	0.625	0.425	0.479	0.375	0.312	0.150	0.312	0.308
CCA	0.633	0.171	0.150	1.000	0.292	0.212	0.188	0.196	0.133	0.337	0.163	0.188	0.171	0.183	0.175	0.142
LDA	0.992	0.508	0.354	0.292	1.000	0.633	0.492	0.462	0.312	0.567	0.371	0.508	0.521	0.362	0.525	0.371
DR-DEA	1.000	0.513	0.267	0.212	0.633	1.000	0.429	0.404	0.275	0.525	0.350	0.492	0.508	0.300	0.517	0.342
BSA	1.000	0.554	0.246	0.188	0.492	0.429	1.000	0.767	0.212	0.763	0.321	0.533	0.558	0.267	0.562	0.313
BCRS	1.000	0.554	0.279	0.196	0.462	0.404	0.767	1.000	0.246	0.742	0.371	0.579	0.592	0.258	0.596	0.329
BI	1.000	0.312	0.625	0.133	0.312	0.275	0.212	0.246	1.000	0.421	0.500	0.321	0.267	0.117	0.267	0.250
AHP-DEA	0.979	0.717	0.425	0.337	0.567	0.525	0.763	0.742	0.421	1.000	0.500	0.704	0.754	0.467	0.750	0.488
NDEA	1.000	0.425	0.479	0.163	0.371	0.350	0.321	0.371	0.500	0.500	1.000	0.517	0.450	0.183	0.450	0.317
EE	1.000	0.637	0.375	0.188	0.508	0.492	0.533	0.579	0.321	0.704	0.517	1.000	0.850	0.400	0.871	0.342
ER	1.000	0.654	0.312	0.171	0.521	0.508	0.558	0.592	0.267	0.754	0.450	0.850	1.000	0.458	0.975	0.342
PEV-PEOI	0.579	0.308	0.150	0.183	0.362	0.300	0.267	0.258	0.117	0.467	0.183	0.400	0.458	1.000	0.467	0.183
NFS-PEOI	1.000	0.654	0.312	0.175	0.525	0.517	0.562	0.596	0.267	0.750	0.450	0.871	0.975	0.467	1.000	0.346
VDMU	1.000	0.367	0.308	0.142	0.371	0.342	0.313	0.329	0.250	0.488	0.317	0.342	0.342	0.183	0.346	1.000

Table 25

Average Normalized Hit Ratio for all pairs of ranking procedures over all artificial datasets with a small number of units K and uniform performance distribution.

	CCR	CE	SE	CCA	LDA	DR-DEA	BSA	BCRS	BI	AHP-DEA	NDEA	EE	ER	PEV-PEOI	NFS-PEOI	VDMU
CCR	1.000	0.164	0.159	0.113	0.292	0.248	0.175	0.164	0.159	0.333	0.184	0.159	0.159	0.089	0.159	0.159
CE	0.164	1.000	0.313	0.143	0.327	0.415	0.522	0.534	0.304	0.508	0.394	0.629	0.646	0.306	0.646	0.360
SE	0.159	0.313	1.000	0.127	0.176	0.178	0.223	0.275	0.625	0.222	0.444	0.375	0.312	0.150	0.312	0.308
CCA	0.113	0.143	0.127	1.000	0.137	0.147	0.147	0.173	0.110	0.158	0.127	0.164	0.148	0.160	0.152	0.118
LDA	0.292	0.327	0.176	0.137	1.000	0.544	0.298	0.287	0.134	0.434	0.177	0.330	0.342	0.184	0.347	0.192
DR-DEA	0.248	0.415	0.178	0.147	0.544	1.000	0.324	0.311	0.186	0.411	0.232	0.403	0.419	0.211	0.428	0.253
BSA	0.175	0.522	0.223	0.147	0.298	0.324	1.000	0.729	0.193	0.545	0.285	0.510	0.539	0.258	0.543	0.295
BCRS	0.164	0.534	0.275	0.173	0.287	0.311	0.729	1.000	0.237	0.543	0.337	0.569	0.585	0.256	0.586	0.318
BI	0.159	0.304	0.625	0.110	0.134	0.186	0.193	0.237	1.000	0.218	0.465	0.321	0.267	0.117	0.267	0.250
AHP-DEA	0.333	0.508	0.222	0.158	0.434	0.411	0.545	0.543	0.218	1.000	0.275	0.501	0.551	0.264	0.547	0.285
NDEA	0.184	0.394	0.444	0.127	0.177	0.232	0.285	0.337	0.465	0.275	1.000	0.488	0.423	0.165	0.423	0.283
EE	0.159	0.629	0.375	0.164	0.330	0.403	0.510	0.569	0.321	0.501	0.488	1.000	0.850	0.400	0.871	0.342
ER	0.159	0.646	0.312	0.148	0.342	0.419	0.539	0.585	0.267	0.551	0.423	0.850	1.000	0.458	0.975	0.342
PEV-PEOI	0.089	0.306	0.150	0.160	0.184	0.211	0.258	0.256	0.117	0.264	0.165	0.400	0.458	1.000	0.467	0.183
NFS-PEOI	0.159	0.646	0.312	0.152	0.347	0.428	0.543	0.586	0.267	0.547	0.423	0.871	0.975	0.467	1.000	0.346
VDMU	0.159	0.360	0.308	0.118	0.192	0.253	0.295	0.318	0.250	0.285	0.283	0.342	0.342	0.183	0.346	1.000

then comment on some interesting aspects observed for *RDM* and *RAM*.

Let us first focus on the pairs of ranking methods that construct the most similar rankings in terms of pairwise preference relations (see Table 26). Specifically, the highest values of Kendall's τ are attained by the pairs of simulation-based methods (for NFS-PEOI and ER – 0.979, EE and ER – 0.927, and NFS-PEOI and EE – 0.926), benchmarking approaches (for BSA and BCRS – 0.851) as well NDEA and SE (0.815). Also, the three methods that compute expected ranks and efficiencies or net flow scores attain similar results to CE, which is understandable in view of related rules that drive their ranking construction procedures. Note that when Kendall's τ is greater than 0.7, this means that the same pairwise relations are observed for over 85% of DMUs. In general, EE, NFS-PEOI, ER, CE, and BCRS attain the highest average similarity with the remaining methods.

On the contrary, the least similarities are observed for comparing PEV-PEOI and CCA with other procedures. Indeed, for all or the vast majority of methods compared with PEV-PEOI and CCA, Kendall's τ is lesser than zero. This means that PEV-PEOI and CCA rank the majority pairs in a different way than other approaches. Another method attaining relatively low similarities when compared with the remaining methods is VDMU. Specifically, the greatest similarity (0.509) that it reaches is when being compared with CE. When comparing the ranking provided by CCR and other procedures, the values of Kendall's τ are relatively high (from 0.311 to 0.670, when neglecting PEV-PEOI and CCA). However, this is mainly due to greater consistency in terms of ordering the inefficient units. Placing all efficient DMUs at the top by CCR deteriorates the similarity measure when compared with the methods imposing a complete order without tied ranks in the vast majority of scenarios.

The observations derived from the analysis of *RDM* and *RAM* in terms of the most and the least similar pairs of methods are very alike (see Tables 27 and 28). In what follows, we raise some interesting points:

- for the vast majority of methods compared with themselves, *RDM* is lower than one, which is due to the shared ranks attained by multiple DMUs in numerous scenarios (e.g., $RDM(CCR, CCR) = 0.729$);
- the greatest similarity in terms of *RDM* is observed for NFS-PEOI and ER, revealing over 98% consistency in terms of the differences between ranks attained by all DMUs;
- the least similarity in terms of *RDM* is observed for BSA and CCA with only 32% consistency (when comparing CCA with all methods, the *RDM* similarity is never higher than 0.35);
- the consistency in terms of rank differences is relatively low, around 50%, for all pairs including AHP-DEA, PEV-PEOI, LDA, and VDMU;

- even though the values of Kendall's τ and *RDM* indicate high similarities, *RAM* confirms that there is no perfect consistency in terms of the ranks attained by the same DMUs for various procedures;
- when analyzing the similarity of rankings in terms of the share of alternatives attaining the same positions, the greatest values are observed for NFS-PEOI and ER as well as BSA and BCRS, pointing out, respectively, over 84% and 76% consistency; these two pairs of procedures along with EE and CCR are also, on average, the most similar in terms of *RAM* to the remaining approaches;
- when comparing CCA and VDMU with other ranking methods, the consistency quantified with *RAM* is always lower than 11% and 19%, respectively.

The detailed results obtained for larger problem instances and performances generated from a normal rather than uniform distribution are discussed in the e-Appendix. In the main paper, we summarize only the main findings. Above all, the conclusions on the sub-groups of ranking methods providing the most and the least similar results are the same irrespective of the considered measure, problem size, and performance distribution. Moreover, when more units are considered, similarity measures for choice get slightly lower, whereas they get higher for ranking. Since when more units are considered, a lesser share of DMUs are tied for the same ranks, NHR is closer to HR and for *RAM* – an even greater polarization of values can be observed (e.g., the most similar methods rank an even greater share of units at the same positions). When input/output performances are normally distributed, the rank similarity measures are slightly higher. Finally, the least stable results for different settings can be observed for methods such as AHP-DEA, CCA, and NFS-PEOI. This can be attributed to the sensitiveness of the eigenvector computations that are involved in these approaches.

4.3. Experimental comparison involving real-world datasets

To confirm the conclusions derived from the analysis based on the artificially generated datasets, we considered ten real-world datasets. In this way, we can verify if the outcomes discussed in Section 4.2 are not affected by the arbitrariness of the generation procedure. The considered datasets represent the most common application areas of the DEA methods, such as finances, education, transportation, healthcare, farming, and the energy industry. These sets involve from 13 to 42 DMUs described in terms of 2–3 inputs and 1–5 outputs (see Table 29).

4.3.1. Measures for choice

The values of measures quantifying the agreement between different

Table 26
Average Kendall's τ for all pairs of ranking procedures over all artificial datasets with a small number of units K and uniform performance distribution.

	CCR	CE	SE	CCA	LDA	DR-DEA	BSA	BCRS	BI	AHP-DEA	NDEA	EE	ER	PEV-PEOI	NFS-PEOI	VDMU
CCR	1.000	0.475	0.602	0.005	0.563	0.561	0.605	0.649	0.451	0.670	0.617	0.462	0.433	-0.004	0.432	0.311
CE	0.475	1.000	0.595	-0.008	0.445	0.579	0.527	0.588	0.586	0.438	0.600	0.737	0.711	-0.008	0.712	0.509
SE	0.602	0.595	1.000	-0.034	0.441	0.502	0.614	0.699	0.692	0.394	0.815	0.601	0.555	-0.016	0.552	0.396
CCA	0.005	-0.008	-0.034	1.000	0.008	-0.002	-0.014	-0.018	-0.028	0.016	-0.026	-0.035	-0.035	-0.017	-0.033	0.003
LDA	0.563	0.445	0.441	0.008	1.000	0.644	0.406	0.451	0.397	0.476	0.439	0.444	0.428	-0.011	0.427	0.288
DR-DEA	0.561	0.579	0.502	-0.002	0.644	1.000	0.474	0.518	0.465	0.480	0.509	0.575	0.557	-0.007	0.557	0.383
BSA	0.605	0.527	0.614	-0.014	0.406	0.474	1.000	0.851	0.454	0.453	0.618	0.515	0.492	-0.031	0.492	0.368
BCRS	0.649	0.588	0.699	-0.018	0.451	0.518	0.851	1.000	0.521	0.473	0.694	0.575	0.545	-0.033	0.544	0.390
BI	0.451	0.586	0.692	-0.028	0.397	0.465	0.454	0.521	1.000	0.381	0.655	0.662	0.627	-0.032	0.627	0.422
AHP-DEA	0.670	0.438	0.394	0.016	0.476	0.480	0.453	0.473	0.381	1.000	0.408	0.442	0.444	-0.029	0.444	0.334
NDEA	0.617	0.600	0.815	-0.026	0.439	0.509	0.618	0.694	0.655	0.408	1.000	0.634	0.590	-0.025	0.590	0.388
EE	0.462	0.737	0.601	-0.035	0.444	0.575	0.515	0.575	0.662	0.442	0.634	1.000	0.927	-0.028	0.926	0.491
ER	0.433	0.711	0.555	-0.035	0.428	0.557	0.492	0.545	0.627	0.444	0.590	0.927	1.000	-0.036	0.979	0.490
PEV-PEOI	-0.004	-0.008	-0.016	-0.017	-0.011	-0.007	-0.031	-0.033	-0.032	-0.029	-0.025	-0.028	-0.086	1.000	-0.034	-0.015
NFS-PEOI	0.432	0.712	0.552	-0.033	0.427	0.557	0.492	0.544	0.627	0.444	0.590	0.926	0.979	-0.034	1.000	0.492
VDMU	0.311	0.509	0.396	0.003	0.288	0.383	0.368	0.390	0.422	0.334	0.388	0.491	0.490	-0.015	0.492	1.000

procedures in terms of indicating the same most preferred DMU are presented in Tables 30 and 31. The analysis of HR confirms that 13 out of 15 ranking methods rank some efficient unit at the top for all ten analyzed datasets (see $HR(CCR, \cdot) = 1$). The two exceptions in this regard are CCA and PEV-PEOI, which rank some inefficient DMU first for 2 out of 10 datasets. Nonetheless, the NHR values are significantly lower (at most 0.271), suggesting that CCR identifies multiple DMUs as efficient. In contrast, other methods indicate only one or a few of them as the most preferred ones.

When it comes to pairs of ranking procedures that attain the greatest similarity in terms of HR and NHR, they partially confirm findings from the analysis of artificial datasets. In particular, a very high similarity is observed for AHP-DEA and CE (HR and NHR equal to, respectively, 0.8 and 0.708), as they rank the same DMU at the very top for 8 out of 10 considered problems. Unlike for the artificially generated datasets, the most preferred DMU indicated by VDMU is often the same as for EE ($HR = NHR = 0.8$), NFS-PEOI ($HR = NHR = 0.8$), and ER ($HR = NHR = 0.7$). Other pairs with relatively high similarity scores in terms of choice include: (EE, ER) and (BI, SE) (for both pairs, $HR = NHR = 0.7$) as well as (EE, AHP-DEA) ($HR = 0.7, NHR = 0.608$). The perfect agreement in terms of indicating the most preferred DMU for 6 out of 10 problem can be observed for all pairs concerning SE, NDEA, and PEV-PEOI ($HR = NHR = 0.6$).

As far as HR and NHR confirming low agreement are concerned, the least obtained values for these measures are equal to 0.1. This means that pairs of methods for which such a value is attained, rank the same DMU at the very top only for a single considered problem. In general, the most significant disagreements in indicating the most preferred DMU are observed for the comparison of CCA or LDA with the remaining approaches. For these two procedures based on statistics and a common set of weights, an average similarity with other methods quantified with HR and NHR is lesser than 0.3. Apart from the pairs involving CCA or LDA, poor agreement degrees can be observed for (DR-DEA, BSA) as well as for the comparisons of NFS-PEOI with NDEA, PEV-PEOI, CE, SE, and BI; ER with CE, SE, and PEV-PEOI; or VDMU with NDEA and PEV-PEOI. For all these pairs, the most advantageous DMU indications are different for at least 7 out of 10 considered datasets.

4.3.2. Measures for ranking

Tables 32–34 provide the similarity measures built on the analysis of entire rankings constructed by different procedures. Although the absolute values of these measures are different than for the artificial datasets, the most and the least similar procedures are the same. In particular, the greatest agreement in terms of the entire rankings can be observed for the following pairs of methods:

- (VDMU, EE) with $\tau = 0.950, RDM = 0.956$, and $RAM = 0.623$;
- (BCRS, SE) with $\tau = 0.944, RDM = 0.948$, and $RAM = 0.809$;
- (SE, NDEA) with $\tau = 0.921, RDM = 0.937$, and $RAM = 0.770$;
- (BCRS, NDEA) with $\tau = 0.899, RDM = 0.923$, and $RAM = 0.744$;
- (VDMU, ER) with $\tau = 0.890, RDM = 0.910$, and $RAM = 0.402$;
- (EE, ER) with $\tau = 0.833, RDM = 0.907$, and $RAM = 0.410$;
- (BCRS, BSA) with $\tau = 0.814, RDM = 0.868$, and $RAM = 0.762$.

For the above pairs, the values of similarity measures indicate very high agreements when taking into account three perspectives, i.e., at least 90% of agreement given pairwise preference relations, at least 86% of consistency in terms of rank differences, and at least 62% of compatibility when it comes to the analysis of ranks attained by the same DMUs (except for (VDMU, ER) and (EE, ER)). As far as the rankings obtained with the standard CCR model are concerned, they are the most similar to the orders imposed by SE, BSA, BCRS, and NDEA. However, this is mainly because these procedures focus only on discriminating between the efficient units while adopting the ranking of inefficient ones after CCR.

When analyzing the similarities for all pairs of methods, the most

Table 27

Average Rank Difference Measure for all pairs of ranking procedures over all artificial datasets with a small number of units K and uniform performance distribution.

	CCR	CE	SE	CCA	LDA	DR-DEA	BSA	BCRS	BI	AHP-DEA	NDEA	EE	ER	PEV-PEOI	NFS-PEOI	VDMU
CCR	0.729	0.629	0.729	0.327	0.588	0.615	0.697	0.729	0.616	0.566	0.729	0.621	0.601	0.368	0.601	0.517
CE	0.629	0.999	0.699	0.343	0.609	0.697	0.659	0.704	0.697	0.608	0.705	0.800	0.780	0.407	0.782	0.644
SE	0.729	0.699	1.000	0.322	0.609	0.644	0.728	0.786	0.768	0.579	0.862	0.705	0.672	0.376	0.669	0.570
CCA	0.327	0.343	0.322	0.981	0.332	0.328	0.322	0.337	0.324	0.334	0.326	0.330	0.329	0.335	0.330	0.347
LDA	0.588	0.609	0.609	0.332	0.856	0.671	0.579	0.612	0.578	0.526	0.605	0.609	0.596	0.365	0.597	0.504
DR-DEA	0.615	0.697	0.644	0.328	0.671	0.928	0.622	0.654	0.622	0.560	0.650	0.694	0.680	0.388	0.680	0.564
BSA	0.697	0.659	0.728	0.322	0.579	0.622	0.961	0.860	0.611	0.579	0.729	0.649	0.633	0.352	0.633	0.537
BCRS	0.729	0.704	0.786	0.337	0.612	0.654	0.860	0.966	0.659	0.605	0.782	0.695	0.675	0.373	0.674	0.563
BI	0.616	0.697	0.768	0.324	0.578	0.622	0.611	0.659	1.000	0.569	0.742	0.744	0.722	0.381	0.722	0.586
AHP-DEA	0.566	0.608	0.579	0.334	0.526	0.560	0.579	0.605	0.569	0.717	0.581	0.611	0.613	0.373	0.614	0.532
NDEA	0.729	0.705	0.862	0.326	0.605	0.650	0.729	0.782	0.742	0.581	0.989	0.730	0.700	0.377	0.698	0.562
EE	0.621	0.800	0.705	0.330	0.609	0.694	0.649	0.695	0.744	0.611	0.730	1.000	0.938	0.460	0.938	0.635
ER	0.601	0.780	0.672	0.329	0.596	0.680	0.633	0.675	0.722	0.613	0.700	0.938	1.000	0.466	0.981	0.632
PEV-PEOI	0.368	0.407	0.376	0.335	0.365	0.388	0.352	0.373	0.381	0.373	0.377	0.460	0.466	1.000	0.467	0.361
NFS-PEOI	0.601	0.782	0.669	0.330	0.597	0.680	0.633	0.674	0.722	0.614	0.698	0.938	0.981	0.467	1.000	0.633
VDMU	0.517	0.644	0.570	0.347	0.504	0.564	0.537	0.563	0.586	0.532	0.562	0.635	0.632	0.361	0.633	1.000

Table 28

Average Rank Agreement Measure for all pairs of ranking procedures over all artificial datasets with small size and uniform performance distribution.

	CCR	CE	SE	CCA	LDA	DR-DEA	BSA	BCRS	BI	AHP-DEA	NDEA	EE	ER	PEV-PEOI	NFS-PEOI	VDMU
CCR	1.000	0.196	0.496	0.108	0.276	0.267	0.543	0.550	0.181	0.461	0.517	0.194	0.180	0.122	0.181	0.135
CE	0.196	1.000	0.230	0.088	0.155	0.226	0.231	0.239	0.226	0.205	0.235	0.328	0.291	0.169	0.299	0.175
SE	0.496	0.230	1.000	0.087	0.150	0.187	0.496	0.526	0.284	0.191	0.594	0.243	0.217	0.134	0.217	0.166
CCA	0.108	0.088	0.087	1.000	0.107	0.099	0.093	0.097	0.083	0.108	0.083	0.091	0.087	0.088	0.089	0.087
LDA	0.276	0.155	0.150	0.107	1.000	0.291	0.156	0.168	0.138	0.270	0.150	0.163	0.149	0.104	0.152	0.114
DR-DEA	0.267	0.226	0.187	0.099	0.291	1.000	0.189	0.198	0.179	0.249	0.199	0.229	0.214	0.140	0.213	0.153
BSA	0.543	0.231	0.496	0.093	0.156	0.189	1.000	0.761	0.183	0.266	0.506	0.227	0.218	0.135	0.220	0.138
BCRS	0.550	0.239	0.526	0.097	0.168	0.198	0.761	1.000	0.199	0.266	0.534	0.245	0.232	0.141	0.234	0.148
BI	0.181	0.226	0.284	0.083	0.138	0.179	0.183	0.199	1.000	0.174	0.255	0.268	0.251	0.153	0.252	0.156
AHP-DEA	0.461	0.205	0.191	0.108	0.270	0.249	0.266	0.266	0.174	1.000	0.203	0.220	0.227	0.143	0.227	0.136
NDEA	0.517	0.235	0.594	0.083	0.150	0.199	0.506	0.534	0.255	0.203	1.000	0.264	0.241	0.139	0.240	0.153
EE	0.194	0.328	0.243	0.091	0.163	0.229	0.227	0.245	0.268	0.220	0.264	1.000	0.619	0.328	0.624	0.185
ER	0.180	0.291	0.217	0.087	0.149	0.214	0.218	0.232	0.251	0.227	0.241	0.619	1.000	0.380	0.845	0.180
PEV-PEOI	0.122	0.169	0.134	0.088	0.104	0.140	0.135	0.141	0.153	0.143	0.139	0.328	0.380	1.000	0.383	0.110
NFS-PEOI	0.181	0.299	0.217	0.089	0.152	0.213	0.220	0.234	0.252	0.227	0.240	0.624	0.845	0.383	1.000	0.180
VDMU	0.135	0.175	0.166	0.087	0.114	0.153	0.138	0.148	0.156	0.136	0.153	0.185	0.180	0.110	0.180	1.000

Table 29

Characteristics of ten real-world datasets used in the experimental analysis.

Source	DMUs	No. of inputs	No. of outputs
Al-Shammari (1999)	15 hospitals in Jordan	3	3
Osman et al. (2011)	32 nurses of Intensive Care Unit	3	5
Thanassoulis and Dunstan (1994)	42 schools in the UK	2	2
Cristobal (2011)	13 renewable energy technologies	3	4
Stokes et al. (2007)	34 dairy farms in Pennsylvania	4	3
Malana and Malano (2006)	25 wheat areas in Pakistan and India	3	1
Gutierrez-Nieto et al. (2007)	30 micro-finance institutions in Latin America	2	3
Kao and Hwang (2008)	24 insurance companies in Taiwan	2	2
Valentine and Gray (2001)	32 ports from all over the world	2	2
Köksal and Aksu (2007)	24 travel agencies in Turkey	3	1

representative rankings are delivered by BCRS, SE, NDEA, and EE. This is confirmed by the highest average similarities to all remaining approaches in terms of Kendalls τ , RDM , and RAM . On the contrary, the least representative rankings are constructed by CCA and LDA. Very low similarity measures confirm the uniqueness of the orders they impose. For example, when comparing the rankings obtained with CCA with the orders determined with any other approach, the highest similarities are extremely low (for $\tau - 0.318$, for $RDM - 0.554$, and for $RAM - 0.118$).

Let us emphasize two additional observations. For the real-world datasets, Kendall's τ and RDM are significantly higher than for the small artificial ones. In particular, there are no negative values for Kendall's τ , and the minimal agreement in terms of rank differences is equal to 37%, hence being by 5% greater than for the artificially generated problems. This suggests that the rankings obtained with different methods for the real-world problems are more similar than for the artificial ones. On the other hand, RAM 's analysis confirms that these rankings do differ vastly for many pairs of procedures. Specifically, for almost 75% of pairs of ranking methods, RAM values are not higher than 0.2. This means that only at most 20% of DMUs attain the same position in the compared rankings.

Table 30
Average Hit Ratio for all pairs of ranking procedures over all real-world datasets.

	CCR	CE	SE	CCA	LDA	DR-DEA	BSA	BCRS	BI	AHP-DEA	NDEA	EE	ER	PEV-PEOI	NFS-PEOI	VDMU
CCR	1.000	1.000	1.000	0.800	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.800	1.000
CE	1.000	1.000	0.500	0.200	0.500	0.500	0.400	0.600	0.500	0.800	0.500	0.400	0.300	0.500	0.300	0.400
SE	1.000	0.500	1.000	0.200	0.100	0.500	0.400	0.600	0.700	0.500	0.600	0.400	0.300	0.600	0.300	0.400
CCA	0.800	0.200	0.200	1.000	0.100	0.200	0.300	0.400	0.200	0.400	0.200	0.300	0.400	0.300	0.200	0.200
LDA	1.000	0.500	0.100	0.100	1.000	0.500	0.300	0.200	0.200	0.500	0.300	0.300	0.300	0.200	0.100	0.300
DR-DEA	1.000	0.500	0.500	0.200	0.500	1.000	0.300	0.400	0.600	0.600	0.500	0.500	0.400	0.300	0.400	0.500
BSA	1.000	0.400	0.400	0.300	0.300	0.300	1.000	0.400	0.400	0.500	0.600	0.500	0.600	0.400	0.400	0.500
BCRS	1.000	0.600	0.600	0.400	0.200	0.400	0.400	1.000	0.500	0.700	0.400	0.500	0.400	0.500	0.500	0.600
BI	1.000	0.500	0.700	0.200	0.200	0.600	0.400	0.500	1.000	0.500	0.500	0.400	0.400	0.500	0.300	0.400
AHP-DEA	1.000	0.800	0.500	0.400	0.500	0.600	0.500	0.700	0.500	1.000	0.500	0.700	0.500	0.500	0.500	0.600
NDEA	1.000	0.500	0.600	0.200	0.300	0.500	0.600	0.400	0.500	0.500	1.000	0.400	0.400	0.600	0.200	0.300
EE	1.000	0.400	0.400	0.300	0.300	0.500	0.500	0.500	0.400	0.700	0.400	1.000	0.700	0.400	0.600	0.800
ER	1.000	0.300	0.300	0.400	0.300	0.400	0.600	0.400	0.400	0.500	0.400	0.700	1.000	0.300	0.500	0.700
PEV-PEOI	1.000	0.500	0.600	0.300	0.200	0.300	0.400	0.500	0.500	0.500	0.600	0.400	0.300	1.000	0.200	0.300
NFS-PEOI	0.800	0.300	0.300	0.200	0.100	0.400	0.400	0.500	0.300	0.500	0.200	0.600	0.500	0.200	1.000	0.800
VDMU	1.000	0.400	0.400	0.200	0.300	0.500	0.500	0.600	0.400	0.600	0.300	0.800	0.700	0.300	0.800	1.000

Table 31
Average Normalized Hit Ratio for all pairs of ranking procedures over all real-world datasets.

	CCR	CE	SE	CCA	LDA	DR-DEA	BSA	BCRS	BI	AHP-DEA	NDEA	EE	ER	PEV-PEOI	NFS-PEOI	VDMU
CCR	1.000	0.187	0.187	0.137	0.187	0.187	0.253	0.187	0.207	0.271	0.187	0.187	0.187	0.207	0.142	0.187
CE	0.187	1.000	0.500	0.200	0.500	0.500	0.333	0.600	0.500	0.708	0.500	0.400	0.300	0.500	0.300	0.400
SE	0.187	0.500	1.000	0.200	0.100	0.500	0.333	0.600	0.700	0.408	0.600	0.400	0.300	0.600	0.300	0.400
CCA	0.137	0.200	0.200	1.000	0.100	0.200	0.233	0.400	0.200	0.308	0.200	0.300	0.400	0.300	0.200	0.200
LDA	0.187	0.500	0.100	0.100	1.000	0.500	0.233	0.200	0.200	0.408	0.300	0.300	0.300	0.200	0.100	0.300
DR-DEA	0.187	0.500	0.500	0.200	0.500	1.000	0.233	0.400	0.600	0.508	0.500	0.500	0.400	0.300	0.400	0.500
BSA	0.253	0.333	0.333	0.233	0.233	0.233	1.000	0.333	0.333	0.341	0.533	0.433	0.533	0.333	0.333	0.433
BCRS	0.187	0.600	0.600	0.400	0.200	0.400	0.333	1.000	0.500	0.608	0.400	0.500	0.400	0.500	0.500	0.600
BI	0.207	0.500	0.700	0.200	0.200	0.600	0.333	0.500	1.000	0.408	0.500	0.400	0.400	0.500	0.300	0.400
AHP-DEA	0.271	0.708	0.408	0.308	0.408	0.508	0.341	0.608	0.408	1.000	0.408	0.608	0.408	0.408	0.408	0.508
NDEA	0.187	0.500	0.600	0.200	0.300	0.500	0.533	0.400	0.500	0.408	1.000	0.400	0.400	0.600	0.200	0.300
EE	0.187	0.400	0.400	0.300	0.300	0.500	0.433	0.500	0.400	0.608	0.400	1.000	0.700	0.400	0.600	0.800
ER	0.187	0.300	0.300	0.400	0.300	0.400	0.533	0.400	0.400	0.408	0.400	0.700	1.000	0.300	0.500	0.700
PEV-PEOI	0.207	0.500	0.600	0.300	0.200	0.300	0.333	0.500	0.500	0.408	0.600	0.400	0.300	1.000	0.200	0.300
NFS-PEOI	0.142	0.300	0.300	0.200	0.100	0.400	0.333	0.500	0.300	0.408	0.200	0.600	0.500	0.200	1.000	0.800
VDMU	0.187	0.400	0.400	0.200	0.300	0.500	0.433	0.600	0.400	0.508	0.300	0.800	0.700	0.300	0.800	1.000

Table 32
Average Kendall's τ for all pairs of ranking procedures over all real-world datasets.

	CCR	CE	SE	CCA	LDA	DR-DEA	BSA	BCRS	BI	AHP-DEA	NDEA	EE	ER	PEV-PEOI	NFS-PEOI	VDMU
CCR	1.000	0.653	0.850	0.250	0.426	0.584	0.800	0.860	0.596	0.638	0.842	0.571	0.534	0.415	0.285	0.549
CE	0.653	1.000	0.728	0.318	0.443	0.675	0.621	0.734	0.657	0.651	0.707	0.627	0.575	0.554	0.345	0.608
SE	0.850	0.728	1.000	0.232	0.444	0.643	0.786	0.944	0.670	0.568	0.921	0.593	0.541	0.473	0.308	0.567
CCA	0.250	0.318	0.232	1.000	0.078	0.175	0.247	0.228	0.155	0.271	0.237	0.206	0.201	0.215	0.157	0.215
LDA	0.426	0.443	0.444	0.078	1.000	0.596	0.411	0.419	0.423	0.406	0.419	0.327	0.322	0.235	0.142	0.319
DR-DEA	0.584	0.675	0.643	0.175	0.596	1.000	0.555	0.628	0.674	0.615	0.611	0.609	0.585	0.467	0.270	0.600
BSA	0.800	0.621	0.786	0.247	0.411	0.555	1.000	0.814	0.545	0.537	0.785	0.546	0.537	0.400	0.311	0.534
BCRS	0.860	0.734	0.944	0.228	0.419	0.628	0.814	1.000	0.669	0.605	0.899	0.628	0.575	0.481	0.339	0.603
BI	0.596	0.657	0.670	0.155	0.423	0.674	0.545	0.669	1.000	0.631	0.654	0.680	0.658	0.659	0.361	0.662
AHP-DEA	0.638	0.651	0.568	0.271	0.406	0.615	0.537	0.605	0.631	1.000	0.547	0.564	0.541	0.522	0.310	0.564
NDEA	0.842	0.707	0.921	0.237	0.419	0.611	0.785	0.899	0.654	0.547	1.000	0.591	0.542	0.467	0.305	0.564
EE	0.571	0.627	0.593	0.206	0.327	0.609	0.546	0.628	0.680	0.564	0.591	1.000	0.883	0.616	0.556	0.950
ER	0.534	0.575	0.541	0.201	0.322	0.585	0.537	0.575	0.658	0.541	0.542	0.883	1.000	0.609	0.491	0.890
PEV-PEOI	0.415	0.554	0.473	0.215	0.235	0.467	0.400	0.481	0.659	0.522	0.467	0.616	0.609	1.000	0.346	0.625
NFS-PEOI	0.285	0.345	0.308	0.157	0.142	0.270	0.311	0.339	0.361	0.310	0.305	0.556	0.491	0.346	1.000	0.561
VDMU	0.549	0.608	0.567	0.215	0.319	0.600	0.534	0.603	0.662	0.564	0.564	0.950	0.890	0.625	0.561	1.000

Table 33
Average Rank Difference Measure for all pairs of ranking procedures over all real-world datasets.

	CCR	CE	SE	CCA	LDA	DR-DEA	BSA	BCRS	BI	AHP-DEA	NDEA	EE	ER	PEV-PEOI	NFS-PEOI	VDMU
CCR	0.889	0.748	0.889	0.504	0.590	0.699	0.846	0.889	0.707	0.681	0.889	0.691	0.659	0.587	0.545	0.675
CE	0.748	0.999	0.797	0.554	0.601	0.760	0.725	0.798	0.739	0.742	0.779	0.727	0.695	0.679	0.580	0.718
SE	0.889	0.797	0.991	0.491	0.598	0.731	0.856	0.948	0.761	0.689	0.937	0.703	0.667	0.629	0.558	0.687
CCA	0.504	0.554	0.491	0.999	0.369	0.451	0.483	0.496	0.439	0.501	0.492	0.491	0.477	0.475	0.467	0.497
LDA	0.590	0.601	0.598	0.369	0.999	0.703	0.568	0.588	0.594	0.576	0.574	0.523	0.518	0.469	0.432	0.514
DR-DEA	0.699	0.760	0.731	0.451	0.703	0.999	0.680	0.726	0.761	0.718	0.713	0.723	0.705	0.629	0.552	0.713
BSA	0.846	0.725	0.856	0.483	0.568	0.680	0.984	0.868	0.673	0.651	0.849	0.676	0.657	0.570	0.553	0.665
BCRS	0.889	0.798	0.948	0.496	0.588	0.726	0.868	0.983	0.752	0.703	0.923	0.723	0.684	0.623	0.578	0.706
BI	0.707	0.739	0.761	0.439	0.594	0.761	0.673	0.752	0.995	0.735	0.745	0.761	0.744	0.750	0.602	0.753
AHP-DEA	0.681	0.742	0.689	0.501	0.576	0.718	0.651	0.703	0.735	0.929	0.680	0.680	0.668	0.652	0.550	0.679
NDEA	0.889	0.779	0.937	0.492	0.574	0.713	0.849	0.923	0.745	0.680	0.997	0.700	0.667	0.624	0.555	0.683
EE	0.691	0.727	0.703	0.491	0.523	0.723	0.676	0.723	0.761	0.680	0.700	0.999	0.907	0.731	0.753	0.956
ER	0.659	0.695	0.667	0.477	0.518	0.705	0.657	0.684	0.744	0.668	0.667	0.907	0.993	0.722	0.703	0.910
PEV-PEOI	0.587	0.679	0.629	0.475	0.469	0.629	0.570	0.623	0.750	0.652	0.624	0.731	0.722	0.995	0.583	0.737
NFS-PEOI	0.545	0.580	0.558	0.467	0.432	0.552	0.553	0.578	0.602	0.550	0.555	0.753	0.703	0.583	0.999	0.762
VDMU	0.675	0.718	0.687	0.497	0.514	0.713	0.665	0.706	0.753	0.679	0.683	0.956	0.910	0.737	0.762	0.999

Table 34
Average Rank Agreement Measure for all pairs of ranking procedures over all real-world datasets.

	CCR	CE	SE	CCA	LDA	DR-DEA	BSA	BCRS	BI	AHP-DEA	NDEA	EE	ER	PEV-PEOI	NFS-PEOI	VDMU
CCR	1.000	0.192	0.737	0.057	0.100	0.129	0.720	0.750	0.146	0.229	0.713	0.163	0.122	0.119	0.133	0.138
CE	0.192	1.000	0.251	0.062	0.144	0.182	0.182	0.224	0.144	0.193	0.211	0.118	0.128	0.143	0.106	0.139
SE	0.737	0.251	1.000	0.060	0.116	0.148	0.715	0.809	0.207	0.173	0.770	0.161	0.128	0.160	0.137	0.144
CCA	0.057	0.062	0.060	1.000	0.046	0.097	0.063	0.068	0.043	0.067	0.062	0.086	0.088	0.081	0.063	0.118
LDA	0.100	0.144	0.116	0.046	1.000	0.156	0.086	0.104	0.104	0.110	0.100	0.085	0.075	0.077	0.054	0.075
DR-DEA	0.129	0.182	0.148	0.097	0.156	1.000	0.122	0.137	0.150	0.140	0.136	0.198	0.221	0.141	0.125	0.173
BSA	0.720	0.182	0.715	0.063	0.086	0.122	1.000	0.762	0.135	0.162	0.692	0.170	0.131	0.122	0.143	0.144
BCRS	0.750	0.224	0.809	0.068	0.104	0.137	0.762	1.000	0.169	0.184	0.744	0.183	0.132	0.129	0.159	0.163
BI	0.146	0.144	0.207	0.043	0.104	0.150	0.135	0.169	1.000	0.212	0.179	0.196	0.153	0.241	0.205	0.190
AHP-DEA	0.229	0.193	0.173	0.067	0.110	0.140	0.162	0.184	0.212	1.000	0.155	0.129	0.116	0.146	0.146	0.120
NDEA	0.713	0.211	0.770	0.062	0.100	0.136	0.692	0.744	0.179	0.155	1.000	0.175	0.125	0.157	0.139	0.147
EE	0.163	0.118	0.161	0.086	0.085	0.198	0.170	0.183	0.196	0.129	0.175	1.000	0.410	0.206	0.434	0.623
ER	0.122	0.128	0.128	0.088	0.075	0.221	0.131	0.132	0.153	0.116	0.125	0.410	1.000	0.173	0.253	0.402
PEV-PEOI	0.119	0.143	0.160	0.081	0.077	0.141	0.122	0.129	0.241	0.146	0.157	0.206	0.173	1.000	0.132	0.207
NFS-PEOI	0.133	0.106	0.137	0.063	0.054	0.125	0.143	0.159	0.205	0.146	0.139	0.434	0.253	0.132	1.000	0.492
VDMU	0.138	0.139	0.144	0.118	0.075	0.173	0.144	0.163	0.190	0.120	0.147	0.623	0.402	0.207	0.492	1.000

5. Conclusions

We considered the problem of ranking Decision Making Units that consume multiple inputs and produce multiple outputs. The methods tackling this problem are considered one of the most important extensions of Data Envelopment Analysis, which has been traditionally oriented only toward dividing the units into efficient and inefficient ones. The contribution of this paper is threefold.

First, we reviewed the methods representing different categories, including cross- and super-efficiency, multivariate statistics, decision analysis, benchmarking, virtual DMU, and social networks. Second, we formalized a novel category of ranking methods based on the concept of Robustness Analysis, which is currently composed of four approaches, including two newly proposed in this paper. These approaches exploit a space of feasible input/output weight vectors with the Monte Carlo simulation to derive the expected efficiencies or ranks, or to compute the priorities or net flow scores of DMUs based on the matrix of pairwise efficiency outranking indices. The use of fifteen methods was illustrated on a numerical example. Third, we compared the rankings constructed by these approaches on 960 artificially generated and 10 real-world datasets with different numbers of inputs and outputs. The results were quantified in terms of five measures, including Hit Ratio and

Normalized Hit Ratio, which focus only on the top-ranked DMUs, as well as Kendall’s τ , Rank Difference Measure, and Rank Agreement Measure, capturing the similarity between the entire rankings.

The analysis of experimental results allowed us to identify three groups of ranking procedures that provide consistent results. The first group comprises the cross-efficiency, three methods based on Robustness Analysis computing the expected efficiencies and ranks or net flow scores, and Virtual DMU. These approaches summarize the efficiency results obtained for multiple feasible weight vectors. For cross-efficiency, these are the most favorable weights for each DMU; for VDMU – the vectors obtained from the analysis of the ideal and anti-ideal units, whereas Robustness Analysis exploits all feasible weights. Even if these methods differ in terms of exploiting ordinal or cardinal measures, focusing on pairwise-oriented or one vs. all comparisons, and incorporating fictive DMUs, the joint idea of building on multiple feasible scenarios implies that the constructed rankings are very similar.

The second group is formed by super-efficiency, BSA, BCRS, BI, and Network DEA. Two main reasons are underlying a high similarity between rankings provided by these methods. Specifically, they investigate the changes in the efficiency scores while focusing on the role of different DMUs as benchmarks. This is attained by quantifying the impacts that each DMU has on the efficiency of remaining ones or that

other DMUs have on a given DMU's efficiency. Moreover, apart from BI, they discriminate between the efficient units while deriving the order of inefficient units from the analysis of standard efficiency scores.

The third group contains LDA and DR-DEA. They incorporate some statistical methods oriented toward finding the set of common weights based on the division of DMUs into the subsets of efficient and inefficient units. Three remaining methods do not belong to the above-described groups. Specifically, Canonical Correlation Analysis, AHP-DEA, and PEV-PEOI incorporate some unique concepts, and significant similarities to the remaining approaches could not be found. In particular, CCA ranks all units using just a single weight vector selected to maximize the correlation between the linear combination of the outputs and inputs. This involves the computation of eigenvalues and eigenvectors (Friedman and Sinuany-Stern, 1997). Although applied to the exploitation of some partial efficiency results, similar operations are performed by AHP-DEA and PEV-PEOI. Many works (see, e.g., Kendall, 1975) indicate that the weights or scores derived from such an analysis are very sensitive, changing drastically from eigenvalue to another or with small deviations in the input matrix. In this perspective, the uniqueness of the rankings delivered by these methods should be attributed to the computational technique incorporated in one of their steps. Nonetheless, due to the specificity of cross-efficiency results exploited by AHP-DEA, the DMUs ranked at the top by this approach are slightly closer to the DMUs indicated by the first group's approaches. In turn, the PEV-PEOI method – building on the robust outcomes – tends to provide a similar ranking to the procedures from the first group on real-world datasets.

The comparative analysis shows clearly that the selection of a ranking method has a significant impact on the obtained results. The choice of such a technique for dealing with a particular study can be conducted based on its properties, strengths and weaknesses, intuitiveness of the underlying idea, and suitability for a specific context of the decision. However, it is also possible to incorporate a few ranking procedures. Our analysis allowed us to identify methods that may be deemed representative for a subset of various procedures (e.g., EE or BCRS) and approaches that offer some unique perspective, hence leading to the rankings that are significantly different from the orders imposed by other algorithms.

We envisage the following directions for future research. First, in this paper, we have selected the representative ranking methods in each category. However, as demonstrated by the extensive reviews by Adler et al. (2002), Aldamak and Zolfaghari (2017), Hinojosa et al. (2017), and Hosseinzadeh Lotfi et al. (2013), many other methods can be included in the experimental comparison. Second, we considered the setting without preference information. The comparison could be extended to the scenarios where constraints on the input/output are available and possibly distinguish various loads of such preference statements. Third, other methods exploiting the results of Robustness Analysis can be proposed. In Multiple Criteria Decision Analysis, such approaches have been recently devised (Kadziński and Michalski, 2016), and they can be adapted to the context of DEA.

CRedit authorship contribution statement

Anna Labijak-Kowalska: Conceptualization, Methodology, Software, Investigation, Data curation, Writing - original draft, Visualization. **Miłosz Kadziński:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing - original draft, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

M. Kadziński acknowledges financial support from the Polish National Science Center under the SONATA BIS project (Grant No. DEC-2019/34/E/HS4/00045). A. Labijak-Kowalska acknowledges support by the research funds (SBAD in 2021) of Poznan University of Technology. The authors thank the three anonymous reviewers whose remarks allowed us to improve the paper significantly.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at <https://doi.org/10.1016/j.eswa.2021.114739>.

References

- Abastante, F., Corrente, S., Greco, S., Ishizaka, A., & Lami, I. M. (2019). A new parsimonious AHP methodology: Assigning priorities to many objects by comparing pairwise few reference objects. *Expert Systems with Applications*, 127, 109–120.
- Adler, N., Friedman, L., & Sinuany-Stern, Z. (2002). Review of ranking methods in the Data Envelopment Analysis context. *European Journal of Operational Research*, 140(2), 249–265.
- Al-Shammari, M. (1999). A multi-criteria data envelopment analysis model for measuring the productive efficiency of hospitals. *International Journal of Operations & Production Management*, 19(9), 879–891.
- Aldamak, A., & Zolfaghari, S. (2017). Review of efficiency ranking methods in data envelopment analysis. *Measurement*, 106, 161–172.
- Alirezadeh, M., & Afsharian, M. (2007). A complete ranking of DMUs using restrictions in DEA models. *Applied Mathematics and Computation*, 189(2), 1550–1559.
- Andersen, P., & Petersen, N. C. (1993). A procedure for ranking efficient units in data envelopment analysis. *Management Science*, 39(10), 1261–1264.
- Barron, F., & Barrett, B. (1996). Decision quality using ranked attribute weights. *Management Science*, 42(11), 1515–1523.
- Bracke, S., Radetzky, M., Rosebrock, C., & Ulutas, B. (2019). Efficiency and effectivity of high precision grinding manufacturing processes: An approach based on combined DEA and cluster analyses. *Procedia CIRP*, 79, 292–297.
- Carrillo, M., & Jorge, J. (2016). A multiobjective DEA approach to ranking alternatives. *Expert Systems with Applications*, 50, 130–139.
- Charnes, A., Cooper, W., Lewin, A., & Seiford, L. (1994). *Data envelopment analysis: Theory, methodology, and applications*. Netherlands: Springer.
- Charnes, A., & Cooper, W. W. (1962). Programming with linear fractional functionals. *Naval Research Logistics Quarterly*, 9(3–4), 181–186.
- Charnes, A., Cooper, W. W., Golany, B., Seiford, L., & Stutz, J. (1985). Foundations of data envelopment analysis for Pareto-Koopmans efficient empirical production functions. *Journal of Econometrics*, 30(1–2), 91–107.
- Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research*, 2(6), 429–444.
- Chen, Y. (2004). Ranking efficient units in DEA. *Omega*, 32(3), 213–219.
- Chu, X., Fielding, G., & Lamar, B. (1992). Measuring transit performance using data envelopment analysis. *Transportation Research Part A: Policy and Practice*, 26(3), 223–230.
- Ciomek, K., & Kadziński, M. (2021). Polyrn: A Java library for sampling from the bounded convex polytopes. *SoftwareX*, 13, Article 100659.
- Cook, W., Liang, L., Zha, Y., & Zhu, J. (2009). A modified super-efficiency DEA model for infeasibility. *Journal of the Operational Research Society*, 60(2), 276–281.
- Cooper, W., Seiford, L., & Zhu, J. (2014). *Handbook on data envelopment analysis*. International series in operations research & management science. Springer, New York.
- Cristobal, J. S. (2011). A multi criteria data envelopment analysis model to evaluate the efficiency of the renewable energy technologies. *Renewable Energy*, 36(10), 2742–2746.
- de Lima Silva, D. F., Ferreira, L., & de Almeida-Filho, A. T. (2020). A new preference disaggregation TOPSIS approach applied to sort corporate bonds based on financial statements and expert's assessment. *Expert Systems with Applications*, 152, Article 113369.
- Doyle, J., & Green, R. (1994). Efficiency and cross-efficiency in DEA: Derivations, meanings and uses. *The Journal of the Operational Research Society*, 45(5), 567–578.
- Emrouznejad, A., & Yang, G.-L. (2018). A survey and analysis of the first 40 years of scholarly literature in DEA: 1978–2016. *Socio-Economic Planning Sciences*, 61, 4–8.
- Farrell, M. J. (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society. Series A (General)*, 120(3), 253–290.
- Fasone, V., & Zapata-Aguirre, S. (2016). Measuring business performance in the airport context: a critical review of literature. *International Journal of Productivity and Performance Management*, 65(8), 1137–1158.
- Fiallos, J., Patrick, J., Michalowski, W., & Farion, K. (2017). Using data envelopment analysis for assessing the performance of pediatric emergency department physicians. *Health Care Management Science*, 20(1), 129–140.
- Friedman, L., & Sinuany-Stern, Z. (1997). Scaling units via the canonical correlation analysis in the DEA context. *European Journal of Operational Research*, 100(3), 629–637.

- Gasser, P., Cinelli, M., Labijak, A., Spada, M., Burgherr, P., Kadziński, M., & Stojadinović, B. (2020). Quantifying electricity supply resilience of countries with robust efficiency analysis. *Energies*, 13, 1535.
- Green, R., Doyle, J., & Cook, W. (1996). Preference voting and project ranking using DEA and cross-evaluation. *European Journal of Operational Research*, 90(3), 461–472.
- Guo, M., Liao, X., & Liu, J. (2019). A progressive sorting approach for multiple criteria decision aiding in the presence of non-monotonic preferences. *Expert Systems with Applications*, 123, 1–17.
- Gutierrez-Nieto, B., Serrano-Cinca, C., & Molinero, C. M. (2007). Microfinance institutions and efficiency. *Omega*, 35(2), 131–142.
- Hatami-Marbini, A., Tavana, M., Agrell, P. J., Hosseinzadeh Lotfi, F., & Ghelej Beigi, Z. (2015). A common-weights DEA model for centralized resource reduction and target setting. *Computers & Industrial Engineering*, 79, 195–203.
- Hays, W. L. & Winkler, R. L. (1970). *Statistics: Probability, inference, and decision*. 1. Holt, Rinehart and Winston.
- Hinojosa, M., Lozano, S., Borrero, D., & Marmol, A. (2017). Ranking efficient DMUs using cooperative game theory. *Expert Systems with Applications*, 80, 273–283.
- Hosseinzadeh Lotfi, F., Jahanshahloo, G. R., Khodabakhshi, M., Rostamy-Malkhalifeh, M., Moghaddas, Z. & Vaez-Ghasemi, M. (2013). A review of ranking models in Data Envelopment Analysis. *Journal of Applied Mathematics* 2013.
- Hosseinzadeh Lotfi, F., Noora, A., Jahanshahloo, G., & Reshadi, M. (2011). One DEA ranking method based on applying aggregate units. *Expert Systems with Applications*, 38(10), 13468–13471.
- Hwang, C.-L., & Yoon, K. (1981). Methods for multiple attribute decision making. In *Multiple attribute decision making* (pp. 58–191). Springer.
- Jahanshahloo, G., Hosseinzadeh Lotfi, F., Khanmohammadi, M., Kazemimanes, M., & Rezaie, V. (2010). Ranking of units by positive ideal DMU with common weights. *Expert Systems with Applications*, 37(12), 7483–7488.
- Jahanshahloo, G. R., Junior, H. V., Lotfi, F. H., & Akbarian, D. (2007). A new DEA ranking system based on changing the reference set. *European Journal of Operational Research*, 181(1), 331–337.
- Joro, T. & Korhonen, P. (2015). *Extension of data envelopment analysis with preference information*. International Series in Operations Research & Management Science. Springer, US.
- Kadziński, M., Badura, J., & Figueira, J. (2020). Using a segmenting description approach in multiple criteria decision aiding. *Expert Systems with Applications*, 147, Article 113186.
- Kadziński, M., Labijak, A., & Napieraj, M. (2017). Integrated framework for robustness analysis using ratio-based efficiency model with application to evaluation of Polish airports. *Omega*, 67, 1–18.
- Kadziński, M., & Michalski, M. (2016). Scoring procedures for multiple criteria decision aiding with robust and stochastic ordinal regression. *Computers & Operations Research*, 71, 54–70.
- Kaffash, S., & Marra, M. (2017). Data envelopment analysis in financial services: A citations network analysis of banks, insurance companies and money market funds. *Annals of Operations Research*, 253(1), 307–344.
- Kao, C., & Hwang, S.-N. (2008). Efficiency decomposition in two-stage data envelopment analysis: An application to non-life insurance companies in taiwan. *European Journal of Operational Research*, 185(1), 418–429.
- Kendall, M. (1975). *Multivariate analysis*. New York: Harrier Press.
- Kohl, S., Schoenfelder, J., Fügner, A., & Brunner, J. O. (2019). The use of data envelopment analysis (dea) in healthcare with a focus on hospitals. *Health Care Management Science*, 22(2), 245–286.
- Köksal, C. D., & Aksu, A. A. (2007). Efficiency evaluation of a-group travel agencies with data envelopment analysis (dea): A case study in the antalya region, turkey. *Tourism Management*, 28(3), 830–834.
- Kritikos, M. N. (2017). A full ranking methodology in data envelopment analysis based on a set of dummy decision making units. *Expert Systems with Applications*, 77, 211–225.
- Lahdelma, R., & Salminen, P. (2006). Stochastic multicriteria acceptability analysis using the data envelopment model. *European Journal of Operational Research*, 170(1), 241–252.
- Liu, J. S., & Lu, W.-M. (2010). DEA and ranking with the network-based approach: a case of R&D performance. *Omega*, 38(6), 453–464.
- Lu, W.-M., & Lo, S.-F. (2009). An interactive benchmark model ranking performers – application to financial holding companies. *Mathematical and Computer Modelling*, 49 (1–2), 172–179.
- Malana, N. M., & Malano, H. M. (2006). Benchmarking productive efficiency of selected wheat areas in pakistan and india using data envelopment analysis. *Irrigation and Drainage: The Journal of the International Commission on Irrigation and Drainage*, 55(4), 383–394.
- Mariz, F. B., Almeida, M. R., & Aloise, D. (2018). A review of dynamic data envelopment analysis: State of the art and applications. *International Transactions in Operational Research*, 25(2), 469–505.
- Matsumoto, K., Makridou, G. & Doumpos, M. (2020). Evaluating environmental performance using data envelopment analysis: The case of European countries. *Journal of Cleaner Production*, 122637 (in press).
- Meng, F., Su, B., Thomson, E., Zhou, D., & Zhou, P. (2016). Measuring china's regional energy and carbon emission efficiency with dea models: A survey. *Applied Energy*, 183, 1–21.
- Nazarko, J., & Saparuskas, J. (2014). Application of dea method in efficiency evaluation of public higher education institutions. *Technological and Economic Development of Economy*, 20(1), 25–44.
- Osman, I. H., Berbary, L. N., Sidani, Y., Al-Ayoubi, B., & Emrouznejad, A. (2011). Data envelopment analysis model for the appraisal and relative performance evaluation of nurses at an intensive care unit. *Journal of Medical Systems*, 35(5), 1039–1062.
- Panayides, P. M., Maxoulis, C. N., Wang, T., & Ng, K. Y. A. (2009). A critical analysis of dea applications to seaport economic efficiency measurement. *Transport Reviews*, 29 (2), 183–206.
- Podinovski, V. (2001). DEA models for the explicit maximisation of relative efficiency. *European Journal of Operational Research*, 131(3), 572–586.
- Saaty, T. (1980). *The Analytic Hierarchy Process: Planning, priority setting, resource allocation*. New York, London: McGraw-Hill International Book Co.
- Salo, A., & Punkka, A. (2011). Ranking intervals and dominance relations for ratio-based efficiency analysis. *Management Science*, 57, 200–214.
- Sexton, T. R., Silkman, R. H., & Hogan, A. J. (1986). Data envelopment analysis: Critique and extensions. *New Directions for Program Evaluation*, 1986(32), 73–105.
- Shen, W.-F., Zhang, D.-Q., Liu, W.-B., & Yang, G.-L. (2016). Increasing discrimination of DEA evaluation by utilizing distances to anti-efficient frontiers. *Computers & Operations Research*, 75, 163–173.
- Sinuany-Stern, Z., & Friedman, L. (1998). DEA and the discriminant analysis of ratios for ranking units. *European Journal of Operational Research*, 111(3), 470–478.
- Sinuany-Stern, Z., Mehrez, A., & Barbooy, A. (1994). Academic departments efficiency via DEA. *Computers & Operations Research*, 21(5), 543–556.
- Sinuany-Stern, Z., Mehrez, A., & Hadad, Y. (2000). An AHP/DEA methodology for ranking decision making units. *International Transactions in Operational Research*, 7 (2), 109–124.
- Stokes, J., Tozer, P., & Hyde, J. (2007). Identifying efficient dairy producers using data envelopment analysis. *Journal of Dairy Science*, 90(5), 2555–2562.
- Tervonen, T., & Lahdelma, R. (2007). Implementing stochastic multicriteria acceptability analysis. *European Journal of Operational Research*, 178(2), 500–513.
- Thanassoulis, E. (1999). Data envelopment analysis and its use in banking. *INFORMS Journal on Applied Analytics*, 29(3), 1–13.
- Thanassoulis, E. & Dunstan, P. (1994). Guiding schools to improved performance using data envelopment analysis: An illustration with data from a local education authority. *Journal of The Operational Research Society* 45, 1247–1262.
- Toma, E., Dobre, C., Dona, I., & Cofas, E. (2015). DEA applicability in assessment of agriculture efficiency on areas with similar geographical patterns. *Agriculture and Agricultural Science Procedia*, 6, 704–711.
- Torgersen, A. M., Forsund, F. R., & Kittelsen, S. A. (1996). Slack-adjusted efficiency measures and ranking of efficient units. *Journal of Productivity Analysis*, 7(4), 379–398.
- Valentine, V. & Gray, R. (2001). The measurement of port efficiency using data envelopment analysis. In: *Proceedings of the 9th world conference on transport research*. Seoul, South Korea.
- Wang, D. D. (2020). Ranking multiple-input and multiple-output units: A comparative study of data envelopment analysis and rank aggregation. *Expert Systems with Applications*, 160, Article 113687.
- Wang, Y.-M., & Luo, Y. (2006). DEA efficiency assessment using ideal and anti-ideal decision making units. *Applied Mathematics and Computation*, 173(2), 902–915.
- Wang, Y.-M., Luo, Y., & Lan, Y.-X. (2011). Common weights for fully ranking decision making units by regression analysis. *Expert Systems with Applications*, 38(8), 9122–9128.
- Wang, Y.-M., & Yang, J.-B. (2007). Measuring the performances of decision-making units using interval efficiencies. *Journal of Computational and Applied Mathematics*, 198(1), 253–267.
- Worthington, A. C. (2001). An empirical survey of frontier efficiency measurement techniques in education. *Education Economics*, 9(3), 245–268.
- Zhu, J. (2014). *Quantitative models for performance evaluation and benchmarking*. International series in operations research & management science. Springer, Switzerland.

Publication [P4]

A. Labijak-Kowalska, M. Kadziński, I. Sychała, L. C. Dias, J. Fiallos, J. Patrick, W. Michalowski, and K. Farion. Performance evaluation of emergency department physicians using robust value-based additive efficiency model. *International Transactions in Operational Research*, 30(1):503–544, 2023, DOI: 10.1111/itor.13099.

Number of citations⁴:

- according to Web of Science: 1
- according to Google Scholar: 3

Contribution of the author of this dissertation and four co-authors:

- Anna Labijak-Kowalska
 - Co-authorship of the idea of adaptation of robustness analysis framework for the value-based DEA model,
 - authorship of the mathematical models for the robustness analysis for VDEA model,
 - authorship of the procedure for estimation the acceptability indices using Hit-And-Run algorithm in a context of value-based DEA model,
 - authorship of the proposed measures for multiple scenarios of efficiency analysis,
 - implementation of the software necessary in the study,
 - application of the robustness analysis methods for the case study,
 - analysis and discussion of the results,
 - authorship of the first draft of the manuscript.
- Inga Sychała
 - Co-authorship of the models for finding the representative set of weights based on the outcomes of the robustness analysis,
 - implementation of the procedures for finding the common set of weights.
- Luis C. Dias
 - Co-authorship of the idea underlying the paper consisting in conducting robustness analysis in the context of value-based efficiency analysis,
 - consultations on the proposed value-based concepts and respective mathematical models ,
 - review and editing of the manuscript.

⁴as on May 30, 2023



- Javier Fiallos
 - collection of the data used in the study of evaluating the performance of emergency department physicians,
 - consulting the results obtained in the study
 - review and editing of the manuscript.

- Wojtek Michalowski
 - Supervision of collecting data used in the study of evaluating the performance of emergency department physicians;
 - co-authorship of the idea of using Data Envelopment Analysis in the context of the case study.
 - consulting the results obtained in the study.
 - review and editing of the manuscript.

WILEY

INTERNATIONAL
TRANSACTIONS
IN OPERATIONAL
RESEARCHIntl. Trans. in Op. Res. 30 (2023) 503–544
DOI: 10.1111/itor.13099

Performance evaluation of emergency department physicians using robust value-based additive efficiency model

Anna Labijak-Kowalska^a, Miłosz Kadziński^{a,*} , Inga Spychała^a, Luis C. Dias^{b,c} , Javier Fiallos^d, Jonathan Patrick^e, Wojtek Michalowski^e and Ken Farion^f

^a*Institute of Computing Science, Poznan University of Technology, Piotrowo 2, Poznań 60-965, Poland*

^b*CeBER, Faculty of Economics, University of Coimbra, Av. Dias da Silva n. 165, Coimbra 3004-512, Portugal*

^c*INESC Coimbra, Department of Electrical and Computer Engineering, University of Coimbra, Rua Sílvio Lima Polo II, Coimbra 3030-290, Portugal*

^d*Elizabeth Bruyere Hospital, 43 Bruyere St., Ottawa, ON K1N 5C8, Canada*

^e*Telfer School of Management, University of Ottawa, 55 Laurier Ave. E, Ottawa, ON K1N 6N5, Canada*

^f*Quality and Systems Improvement, Children's Hospital of Eastern Ontario, 401 Smyth Road, Ottawa, ON K1H 8L1, Canada*

*E-mail: anna.labijak@cs.put.poznan.pl [Labijak-Kowalska]; milosz.kadzinski@cs.put.poznan.pl [Kadziński];
ingaspychala@gmail.com [Spychala]; lmcdias@fe.uc.pt [Dias]; javier.fiallos@gmail.com [Fiallos];
patrick@telfer.uottawa.ca [Patrick]; wojtek@telfer.uottawa.ca [Michalowski]; farion@cheo.on.ca [Farion]*

Received 28 October 2020; received in revised form 20 September 2021; accepted 27 November 2021

Abstract

We propose a novel variant of the value-based additive data envelopment analysis model. It conducts a comprehensive robustness analysis of efficiency outcomes for all feasible input and output weights using mathematical programming and the Monte Carlo simulation. We also introduce the original procedures for selecting a common vector of weights and an approach for investigating the stability of results in a multiscenario setting. The presented framework is applied to evaluate the performance of emergency department physicians using data from the Children's Hospital of Eastern Ontario in Ottawa. Our focus is on the physicians' performance when dealing with groups of patients' complaints related to abdominal pain and constipation, fever, extremity injury, head injury, and laceration/puncture. The obtained results emphasize the strong dependence of the physicians' performances on the selected weight vectors. However, they prove helpful in pointing out overall good performers who can serve as universal benchmarks or niche performers being markedly better in providing care to a given complaint group. They also offer a basis for developing an improvement plan for the underperforming physicians, identifying the priorities for a practice-oriented model, and recognizing the most challenging patients' complaints.

Keywords: data envelopment analysis; physician's performance; emergency department; robustness analysis; efficiency analysis; value-based additive efficiency; multiattribute value function; common set of weights

*Corresponding author.

© 2021 The Authors.

International Transactions in Operational Research © 2021 International Federation of Operational Research Societies.

Published by John Wiley & Sons Ltd, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main St, Malden, MA02148, USA.

1. Introduction

Measuring the performance in healthcare is a complex, multidimensional problem. At each level—from individual physicians through medical practice and tertiary care to the entire healthcare system—one expects that the properly working unit or institution provides the best possible care efficiently using available resources. The variety of indicators does not allow for a direct performance assessment by monitoring only selected individual measures that should be optimized. In turn, it is required to find a proper trade-off between consumed resources and the quality of provided care.

One of the primary methods for assessing healthcare efficiency is through patient satisfaction surveys, using, for example, a predefined Likert scale (Smith et al., 2004; Jennings et al., 2009). However, such a survey-based approach gives information only about the patients' perceptions while failing to capture how efficiently resources are utilized. Another approach that, in turn, considers multiple performance aspects consists of using the composite indicators to aggregate all the individual indicators into a single quality measure (Goddard and Jacobs, 2009). Nonetheless, this method requires selecting an appropriate aggregation approach and an arbitrary parametrization with weights associated with different indicators. A slight change in these subjective values may vastly influence the relative performance evaluation of healthcare units (Jacobs et al., 2005).

The subjectivity and arbitrariness issues related to setting the weight values are no longer present when using the data envelopment analysis (DEA) (Charnes et al., 1978), that is, a nonparametric efficiency evaluation method. This approach allows measuring the relative efficiency of decision-making units (DMUs), which consume multiple inputs (resources) and produce multiple outputs (effects). DEA allows performing evaluation and measurement without assigning prior weights. In turn, one DMU's efficiency score depends on the input and output values of others. These aspects contribute to DEA's applicability, making the results objective in relation to the scores computed using composite indicators.

Healthcare is, next to banking, agriculture, transportation, and education, one of the most common application areas of DEA (Liu et al., 2013). The most frequently considered DMUs are hospitals. Kohl et al. (2019) provided an in-depth review of DEA applications in healthcare with a particular focus on hospitals. Recent examples include evaluating the Greek National Health Service hospitals (Flokou et al., 2017), investigating an impact of the economic recession on the performance of hospitals in Pennsylvania (Chen et al., 2019), or assessment of the technical efficiency of a few hundred Turkish hospitals (Küçük et al., 2020).

When it comes to other types of DMUs, medical-group practices are becoming increasingly popular. Andes et al. (2002) investigated the organizational factors affecting the overall physician practice efficiency for over one hundred primary care physician practices in the United States. Furthermore, Testi et al. (2013) assessed the primary care physician practices in Italy when treating diabetic patients. In Portugal, primary healthcare units were assessed from a perspective of geographical inequity (Amado and Santos, 2009) and comparing two types of units (Gouveia et al., 2016). DEA has also been used to evaluate individual departments, such as emergency departments (EDs) or operating rooms (ORs). In particular, Kang et al. (2017) examined the efficiency of EDs to help hospitals plan the redesign. Ketabi et al. (2018) and Akkan et al. (2020) evaluated EDs of hospitals in Isfahan (Iran) and Istanbul (Turkey), intending to identify the improvement strategies for

the underperforming units. In turn, Basson and Butler (2006) compared multiple ORs to propose new resource allocations to improve their performance.

DEA has also proved helpful in evaluating the performance of nursing homes. In this context, a combination of different DEA models was used to study the care planning process, determine the best techniques to ensure the quality of care, and identify the determinants that affect the homes' efficiency. Such studies were conducted for the nursing homes in The Netherlands (Lee et al., 2009), Portugal (Veloso et al., 2018), and the United States (Kooreman, 1994; Shimshak et al., 2009). Other types of healthcare institutions evaluated included fire and emergency services (Choi, 2005), visiting nurse service agencies (Kuwahara et al., 2013), and health maintenance organizations (Siddharthan et al., 2000). DEA was also used to evaluate the performance of national healthcare systems (Zehra and Serpil, 2018).

Finally, DEA was used for the assessment of individual physicians working in the hospital. Chilingerian and Sherman (1990) identified the inefficient practice patterns of the physicians treating cardiac patients, whereas Wagner and Shimshak (2000) evaluated the primary care physicians from a managed care organization. Furthermore, Ozcan et al. (2000) compared the resource utilization between medical specialists in the treatment of Medicaid sinusitis patients in Virginia. Also, Johannessen et al. (2017) investigated the impact of hospital reform in Norway on the performance of individual physicians. Finally, Fiallos et al. (2017) developed a model to assess ED physicians' performance taking into account different complaint groups and different types of medical trainees.

The literature on the DEA-based evaluation of healthcare units is rich. Both standard (Basson and Butler, 2006; Kuwahara et al., 2013) and enhanced DEA models (e.g., network DEA; Khushalani and Ozcan, 2017; Gerami et al., 2020 or window-DEA; Flokou et al., 2017) have been used. Moreover, DEA has also been combined with statistical analysis (Chilingerian, 1995; Akkan et al., 2020), multiple criteria decision analysis (MCDA); Rouyendegh et al., 2019), or machine learning (Tosun, 2012). All these applications assess each DMU based on a single vector of input/output weights, namely the vector that yields the most favorable assessment for that unit. Yet, as the choice of any specific vector is open to debate, it is worth analyzing how assessments would change when applying other feasible weight vectors. A noteworthy exception in this regard is the work of Schang et al. (2016) who used ratio-based efficiency analysis (Salo and Punkka, 2011) to evaluate the impact of the chosen weights on the final score of composite indicators applied for evaluating a set of Scottish Health Boards.

This paper introduces a novel robust value-based framework for efficiency analysis. Specifically, we extend the value-based additive DEA (VDEA) model (Gouveia et al., 2008), which combines DEA with the multiattribute value theory (MAVT) (Keeney and Raiffa, 1993). The underlying idea is to convert the relevant inputs and outputs into criteria associated with marginal value functions and aggregate them using an additive model. In the standard VDEA model, each DMU can choose the weights associated with the marginal value functions that minimize the difference of comprehensive value (efficiency score) to the best DMU. In turn, we investigate the robustness of results attained for all feasible input and output weights. We deliver the outcomes referring to the efficiency measures, ranks, and preference relations, and for each of those, we propose methods based on mathematical programming providing information about the extreme values (minimum and maximum) obtained for a given result. As the differences between the extreme bounds are often large, the robustness analysis framework also incorporates stochastic methods based on the Monte

Carlo simulations. They are useful for estimating the distributions of the considered measures or relations. Such distributions are captured by acceptability indices, quantifying the proportion of feasible weights confirming a given result. For the purpose of this work, these methods have been implemented and made available as independent modules on the open-source *diviz* platform (Meyer and Bigaret, 2012).

The proposed methodological framework is also enriched in two ways. On the one hand, we introduce novel procedures for computing a representative vector of common weights that allows ranking all DMUs univocally. Such a vector is chosen to match as well as possible the conclusions obtained through the robustness analysis. In case one unit is robustly better than the other, the selected vector should emphasize this advantage. In turn, for DMU pairs that are indistinguishable in terms of robust results, the chosen vector should make the difference between these DMUs as small as possible. On the other hand, we provide methods for quantifying the results' robustness under different evaluation scenarios. These outcomes consider two levels of robustness. The first level refers to the robustness of outcomes for an individual scenario, whereas the second captures the stability of results given the multiplicity of possible scenarios.

We applied the proposed robust value-based efficiency analysis methods for evaluating the performance of ED physicians using data from the Children's Hospital of Eastern Ontario (CHEO) in Ottawa, Canada. We consider three inputs (the average encounter time per patient visit, the average number of laboratory tests per patient visit, and the average number of radiology orders per patient visit), and one output (rate of nonreturn patient visits within 72 hours). Our primary focus is on a group of patients with primary complaints upon presentation being abdominal pain and constipation. However, in a multiscenario analysis, we also consider two other complaint groups related to fever and lower or upper extremity injury, head injury, and laceration/puncture.

In Fiallos et al. (2017), the performance of the same physicians was evaluated using an original SBM-SWAT VRS efficiency model. The main motivation for its use was to penalize a “compensatory behavior,” that is, preventing some physicians from being judged as efficient because of attaining advantageous results on only a single input or output. However, such an approach considers an extremely limited space of symmetric weights, hence clearly favoring the physicians with balanced performance profiles. Also, it involves an arbitrary parameterization of the model with precise values of symmetry factors (β) that are difficult to specify or determine experimentally. Finally, it derives the efficiency measures from analyzing the most favorable weight vector for each physician, providing precise scores that do not offer a common basis for physicians' comparison.

We demonstrate that a more fair and justified way for preventing the above-mentioned “compensatory effect” is to conduct a robustness analysis. It provides meaningful means for comparing physicians based on their performance for diverse scenarios relevant to efficiency analysis. The robust results are less affected by the inclusion or removal of a single physician; they can be derived when the number of DMUs is relatively small compared to the number of inputs and outputs, while highly discriminating between physicians. In this way, we may identify the subsets of the most distinguishing and underperforming physicians while counteracting the “compensatory effect” related to excelling at only a single aspect of their clinical role and performing poorly on the remaining inputs and outputs. Moreover, we demonstrate that the application of our framework is beneficial for directly comparing pairs of physicians, yielding insights for identifying potential outliers, and proposing gradual improvement paths for the DMUs.

The remainder of this paper is organized as follows. Section 2 describes the novel robust value-based methods for efficiency analysis. In Section 3, we describe a case study. Section 4 concludes the paper and discusses the potential implications of the proposed approach.

2. Robustness analysis framework for value-based additive efficiency analysis

In this section, we present the robustness analysis methods for the value-based additive efficiency model. First, we will remind the basic framework that selects for each DMU the input and output weights that minimize the difference of comprehensive value to the best DMU. We will then discuss two streams of procedures for investigating the robustness of efficiency results attained for all feasible input and output weights. Moreover, we will generalize the proposed approaches for investigating the stability of results in multiple scenarios that can be considered for the same DMUs. Finally, we will present the algorithms for selecting a common vector of weights based on robust outcomes.

In what follows, we will use the following notation:

- K —a number of units (DMUs);
- \mathcal{D} —a finite set of DMUs, $\mathcal{D} = \{\text{DMU}_1, \dots, \text{DMU}_K\}$;
- N and M —the number of inputs and outputs, respectively;
- $Q = N + M$ —a number of all factors relevant for the analysis;
- w_q —a weight associated with the q th factor (input or output);
- u_q —a marginal value function associated with the q th factor;
- $S_w = \{w = (w_1, w_2, \dots, w_q)^T \mid w \geq 0, A_w w \leq 0\}$ —a space of feasible weight vectors, where A_w is the coefficient matrix of user-defined linear weight constraints.

2.1. Reminder on value-based additive data envelopment analysis

DEA encompasses several models that can be used to measure the relative efficiency of DMUs. In the most standard approach, the efficiency is expressed as a ratio between a single virtual output and a single virtual input, that is, weighted sums of outputs and inputs, respectively (Charnes et al., 1978). The seminal CCR (Charnes et al., 1978) and BCC models (Banker et al., 1984) belong to this category of radial models, in which the weights involved in the efficiency measure are established by identifying the most advantageous scenario for the DMU under evaluation. Later, several nonradial models have been proposed, such as the directional distance function (Färe and Grosskopf, 2000) and the additive model (Charnes et al., 1985). All these methods share the core DEA features of considering an empirical production possibility set and allowing each DMU under evaluation to select the weights involved in the definition of efficiency in a way that makes its efficiency score as good as possible. When using an additive efficiency model (Charnes et al., 1985), the underlying idea is to maximize the L_1 distance of each DMU to the efficient frontier. A few issues can be associated with this model: the comparability of the scales on which the inputs and outputs are expressed, the very pessimistic character of the derived efficiency

measures, and the lack of their intuitive interpretation. To address these issues, Gouveia et al. (2008) proposed a variant of an additive DEA model, exploiting the links between DEA and MAVT. In this approach, the DMUs are treated as decision alternatives evaluated in terms of multiple relevant criteria. Each criterion corresponds to an input or an output factor in the traditional efficiency model. Specifically, a comprehensive value E_o of DMU_o is computed using an additive value function, that is, a weighted sum of the marginal values assigned to the performance on each factor:

$$E_o = \sum_{q=1}^Q w_q u_q(DMU_o), \quad (1)$$

where w_q is the weight, interpreted as a scale coefficient of the marginal value functions u_j , such that $w_q, q = 1, \dots, Q$, and $\sum_{q=1}^Q w_q = 1$. Moreover, a preference direction is associated with each factor $q, q = 1, \dots, Q$. Function u_j takes values between 0 and 1, being nonincreasing for the criteria corresponding to inputs and nondecreasing for outputs. In this way, lesser inputs and greater outputs are more preferred and all inputs and outputs are expressed in comparable value scales. Overall, the comprehensive value lies in the range of $[0, 1]$. Using the above model, the efficiency of DMU_o relatively to the set of DMUs can be verified by solving the following linear programming (LP) problem:

$$\text{Minimize } d_o \quad (2)$$

s.t.

$$\left. \begin{array}{l} \sum_{q=1}^Q w_q u_q(DMU_k) - \sum_{q=1}^Q w_q u_q(DMU_o) \leq d_o, \text{ for } k = 1, \dots, K, \\ d_o \geq 0, \\ \sum_{q=1}^Q w_q = 1, \\ w_q \geq 0, \quad q = 1, \dots, Q, \\ \mathbf{w} \in S_w. \end{array} \right\} \mathcal{W}.$$

This LP minimizes the distance d_o of analyzed $DMU_o \in \mathcal{D}$ to the unit with the greatest comprehensive value. If the least distance ($d_{*,o}$) is equal to 0, then DMU_o is considered efficient. It means that there exists some feasible weight vector for which DMU_o attains a comprehensive value not worse than the value of all other units. Otherwise, that is, if $d_{*,o} > 0$, DMU_o is not efficient, and $d_{*,o}$ reflects a “min-max regret” perspective. In the following sections, we will denote a set of constraints specifying all feasible, nonnegative, and normalized weights by \mathcal{W} .

The assessment of a DMU_o with E_o and $d_{*,o}$ reflects two different perspectives. On the one hand, E_o might be called an absolute efficiency score, as it is independent of the other DMUs. It indicates a score in $[0, 1]$, where 1 corresponds to an ideal situation in which a DMU has a value of 1 on

every criterion, that is, it produces the maximum amount of outputs with the minimum amount of inputs, whereas 0 corresponds to a value of 0 on every criterion, that is, it produces the minimum amount of outputs with the maximum amount of inputs. On the other hand, $d_{*,o}$ corresponds to a DEA relative efficiency, that is, relative to the empirically observed efficient frontier, which could change if other DMUs were added or excluded from \mathcal{D} .

Let us emphasize that in terms of MCDA, DMUs with $d_{*,o} = 0$ would be formally called “weakly efficient.” It is possible that a dominated unit would attain a comprehensive value that is at least as good as all other units’ scores. If this effect was undesired, one could either assume that the weights w_q , $q = 1, \dots, Q$, should be positive or solve a second LP problem to maximize the minimal weight values for $d_o = 0$ (Gouveia et al., 2008).

When computing $d_{*,o}$, only the input/output weight vector most favorable to DMU_o is taken into account, which limits the insights that can be obtained from the analysis. First, it makes the comparison of efficiency scores questionable due to the nonuniqueness of the weight vectors favorable to each DMU, that is, the analyst lacks a common basis to analyze the attained efficiencies. Second, such an analysis neglects other weight vectors that could provide a realistic setting for the comparison of DMUs, potentially leading to useful information on the variety of efficiency scores under a variety of scenarios. Third, it provides limited means for discriminating between the units. This is particularly true when the number of considered factors is large, implying that a large subset of DMUs can be deemed efficient.

The limitations of using a single weight vector motivated the development of methods for robustness analysis (Lahdelma and Salminen, 2006; Salo and Punkka, 2011; Kadziński et al., 2017). Their essence consists of investigating the stability of outcomes for all feasible weights associated with the inputs and outputs. In what follows, we discuss the methods that incorporate the mathematical programming techniques to capture the exact, extreme outcomes, or the Monte Carlo simulation to estimate the distribution of results observed for feasible weights. When doing so, we assume a uniform distribution of weights. In this way, each weight vector has equal chances ($= 1/\text{vol}(W)$, where $\text{vol}(W)$ is the volume of the feasible weight space) to be considered within a sample of weights derived in the simulation. However, it is also possible to use the method with some exogenously given weight distribution.

2.2. Robustness analysis with mathematical programming

In this section, we discuss the mathematical models for computing the extreme efficiency results observed in the set of all feasible weights. We refer to three types of outcomes: efficiency scores, ranks, and pairwise preference relations.

When it comes to the efficiency scores, we may consider the relative distances or absolute values. For the former (Gouveia et al., 2008), we are interested in the range $[d_{*,o}, d_o^*]$ delimited by the least $d_{*,o}$ and the greatest d_o^* possible distance of DMU_o from the efficient unit that attains the maximal comprehensive value for a given weight vector. The minimal distance $d_{*,o}$ can be computed as explained in Section 2.1, whereas the maximal one, d_o^* , can be obtained by solving the following mixed-integer linear programming (MILP) model:

$$\text{Maximize } d_o \tag{3}$$

s.t.

$$\left. \begin{aligned} \sum_{q=1}^Q w_q u_q(\text{DMU}_k) - d_o &\geq \sum_{q=1}^Q w_q u_q(\text{DMU}_o) - C(1 - b_k), \text{ for } k = 1, \dots, K, k \neq o, \\ \sum_{k=1, \dots, K; k \neq o} b_k &= 1, \\ b_k &\in \{0, 1\}, \text{ for } k = 1, 2, \dots, K; k \neq o, \\ d_o &\geq 0, \\ \mathcal{W}, \end{aligned} \right\}$$

where C is a large positive constant. The above model maximizes the distance of DMU_o from some other DMU. The first four constraints guarantee that d_o is equal to the difference between E_k and E_o for some $k \in \{1, \dots, K\}$ and $k \neq o$. Note that if a binary variable $b_k \in \{0, 1\}$ is equal to 0, then the first constraint is always satisfied, whereas in case $b_k = 1$, then $C(1 - b_k) = 0$ and $d_o = E_k - E_o$. We require that the latter holds for some $\text{DMU}_k, k \in \{1, \dots, K\}$ and $k \neq o$.

In turn, the interval $[E_{*,o}, E_o^*]$, delimited by the least $E_{*,o}$ and the greatest E_o^* efficiency scores, can be determined by optimizing the comprehensive value of DMU_o subject to the constraints defining a set of admissible inputs and output weights, that is,

$$\text{Minimize/maximize } \sum_{q=1}^Q w_q u_q(\text{DMU}_o), \text{ s.t. } \mathcal{W}. \tag{4}$$

The rank-oriented perspective offers greater stability because it is based on ordinal rather than cardinal comparisons (Salo and Punkka, 2011). Note that some small changes in the data that might change DMU scores might still keep the ranking of the DMUs unchanged (Kadziński et al., 2017). To compute the best (minimal) possible $R_{*,o}$ rank for DMU_i , the following MILP problem needs to be solved:

$$\text{Minimize } 1 + \sum_{k=1, \dots, K; k \neq o} b_k \tag{5}$$

s.t.

$$\left. \begin{aligned} \sum_{q=1}^Q w_q u_q(\text{DMU}_k) - \sum_{q=1}^Q w_q u_q(\text{DMU}_o) &\leq C b_k, \text{ for } k = 1, \dots, K; k \neq o, \\ b_k &\in \{0, 1\}, \text{ for } k = 1, 2, \dots, K; k \neq o, \\ \mathcal{W}. \end{aligned} \right\}$$

The above problem minimizes the number of DMUs that, for some feasible weight vector, attain greater (absolute) efficiency than DMU_o . Such a number increased by one is equal to

$R_{*,o}$ (Kadziński et al., 2012b). To compute the worst (maximal) possible R_o^* rank for DMU_o , we need to maximize the number of DMUs with the efficiency scores greater than the efficiency of DMU_o . This can be attained by solving the following MILP problem:

$$\text{Maximize } 1 + \sum_{k=1, \dots, K; k \neq o} b_k$$

s.t.

$$\left. \begin{aligned} \sum_{q=1}^Q w_q u_q(DMU_o) - \sum_{q=1}^Q w_q u_q(DMU_k) &\leq C(1 - b_k), \text{ for } k = 1, \dots, K; k \neq o, \\ b_k &\in \{0, 1\}, \text{ for } k = 1, 2, \dots, K; k \neq o, \\ \mathcal{W}. \end{aligned} \right\} \quad (6)$$

The extreme relative distances, absolute values, and ranks indicate the performance of each DMU in the least and the most favorable scenarios that correspond to the pessimistic and optimistic settings, respectively. When referring to the latter concepts, we will mean the weight vectors for which a DMU attains the worst or the best results in the entire space of feasible input and output weights from a particular outcome perspective.

It is also possible to compare DMUs in a pairwise fashion concerning their efficiencies for all feasible weights. This efficiency-based binary relation, which we call pairwise preference relation, is defined for any pair of DMUs, being independent of the remaining DMUs. Given a set of feasible weights associated with different factors, two certainty levels can be considered (Greco et al., 2008). On the one hand, the possible preference relation \succsim_E^P holds for a pair (DMU_o, DMU_k) if $E_o \geq E_k$ for at least one feasible weight vector. To verify its truth, the following LP model needs to be solved:

$$\text{Maximize } d_{\alpha,k}, \text{ s.t. } \sum_{q=1}^Q w_q u_q(DMU_o) - \sum_{q=1}^Q w_q u_q(DMU_k) \geq d_{\alpha,k} \text{ and } \mathcal{W}. \quad (7)$$

If the maximal attained value of $d_{\alpha,k}$ is not lesser than 0, there exists at least one feasible vector \mathbf{w} for which $E_o \geq E_k$, and thus $DMU_o \succsim_E^P DMU_k$. On the other hand, the necessary preference relation \succsim_E^N holds for a pair (DMU_o, DMU_k) if $E_o \geq E_k$ for all feasible weight vectors. Its truth can be verified by considering the following LP problem:

$$\text{Minimize } d_{\alpha,k}, \text{ s.t. } \sum_{q=1}^Q w_q u_q(DMU_o) - \sum_{q=1}^Q w_q u_q(DMU_k) \leq d_{\alpha,k} \text{ and } \mathcal{W}. \quad (8)$$

When the minimal value of $d_{\alpha,k}$ is greater than or equal to 0, there is no feasible weight vector for which $E_k > E_o$. This, in turn, implies that $E_o \geq E_k$ holds for all feasible weights, and thus $DMU_o \succsim_E^N DMU_k$.

Analyzing exact robust results is fundamental in decision problems with high stakes, where the specification of weight constraints is impossible or hampered by significant uncertainties. Then,

the variety of results is greater, and one may implement more “precautionary” rules based on the analysis of the worst possible results (in case of efficiency scores or ranks) or the necessary outcomes (in case of preference relations).

2.3. Robustness analysis methods for multiple scenarios of efficiency evaluation

The robustness analysis methods for DEA have been initially designed for dealing with a single scenario (Kadziński et al., 2017; Salo and Punkka, 2011), representing a particular evaluation context for a set of homogeneous DMUs. In such a scenario, the units are characterized by precise values of inputs and outputs. However, in some situations, the same set of DMUs could be evaluated under multiple scenarios. Let us denote a set of such scenarios by \mathcal{S} . For each DMU, the input and output values may differ from one scenario to another, hence potentially leading to different efficiency results. For example, in the study discussed in this paper, the scenarios correspond to different complaint groups, with complaints in each group forming a relatively homogenous population regarding ED management. It does not make sense to jointly consider different clinical and diagnostic categories, as this would lead to an averaging effect. Practice variations are expected and observed across presenting complaints due to the difference in resource utilization patterns for each type of complaint. This motivated accounting for each group separately and producing performance evaluations per type of complaint.

In this section, we extend the robust methods to address such multi-scenario settings. This is attained by adopting the approaches proposed initially for dealing with group decision-making problems (Greco et al., 2012). The multiscenario robust results consider two levels of certainty for the efficiency outcomes. The first level refers to the robustness analysis results for each scenario $S \in \mathcal{S}$. In what follows, we focus only on the exact outcomes computed with mathematical programming (see Section 2.2). Let us denote the extreme distances to the efficient unit by $[d_{*,o,S}, d_{o,S}^*]$, the extreme efficiency scores by $[E_{*,o,S}, E_{o,S}^*]$, the extreme ranks by $[R_{*,o,S}, R_{o,S}^*]$, and the necessary and possible preference relations by $\succsim_{E,S}^N$ and $\succsim_{E,S}^P$, respectively. The other level concerns the support given to some robust results by different scenarios. For this purpose, we consider the necessary and possible support depending on whether some outcome is confirmed by all or at least one scenario, respectively. Without loss of generality, we define the considered results only in the context of the necessary preference relation and extreme efficiency scores, and they can be generalized analogously to the possible relation, extreme distances, and scores:

- the necessary-necessary preference relation $\succsim_{E,S}^{N,N}$ holds for $(DMU_o, DMU_k) \in \mathcal{D} \times \mathcal{D}$ if for all $S \in \mathcal{S}$, $DMU_o \succsim_{E,S}^N DMU_k$;
- the necessary-possible preference relation $\succsim_{E,S}^{N,P}$ holds for $(DMU_o, DMU_k) \in \mathcal{D} \times \mathcal{D}$ if for at least one $S \in \mathcal{S}$, $DMU_o \succsim_{E,S}^N DMU_k$;
- the set of possible-necessary efficiency ranks $[R_{*,o,S}^N, R_{o,S}^{*,N}]$ is a set of ranks attained for all $S \in \mathcal{S}$, that is, $[R_{*,o,S}^N, R_{o,S}^{*,N}] = \bigcap_{S \in \mathcal{S}} [R_{*,o,S}, R_{o,S}^*]$;
- the set of possible-possible efficiency ranks $[R_{*,o,S}^P, R_{o,S}^{*,P}]$ is a set of ranks attained for at least one $S \in \mathcal{S}$, that is, $[R_{*,o,S}^P, R_{o,S}^{*,P}] = \bigcup_{S \in \mathcal{S}} [R_{*,o,S}, R_{o,S}^*]$.

Note that in the case of great divergence of results for various scenarios, $[R_{*,o,S}^N, R_{o,S}^{*,N}]$ can be empty, whereas $[R_{*,o,S}^P, R_{o,S}^{*,P}]$ does not need to be continuous, that is, there can be some holes in the range delimited by $R_{*,o,S}^P$ and $R_{o,S}^{*,P}$. In any case, these outcomes are useful for verifying the stability of performance under multiple scenarios, indicating the spaces of agreement and discordance for the same unit or pair of DMUs.

2.4. Robustness analysis with the Monte Carlo simulation

In most decision problems, the difference between the extreme distances, scores, or ranks is large, the possible relation is rich, whereas the necessary one is relatively poor. Thus, it is important to determine the distribution of distances, scores, ranks, and relations over the feasible weight space. Such a probability distribution can be estimated with Monte Carlo simulations. To generate a random sample of weights, we apply the hit-and-run algorithm (Ciomek and Kadziński, 2021). In general, it is possible to use any arbitrarily chosen probability distribution on the joint density function in the feasible weight space. When it can be reliably defined, the evaluation model reflects the DM's preferences more faithfully. However, elicitation of a fully specified probability distribution calls for a major effort. When it is not possible, a standard assumption—also made in this paper—is to consider weights that are uniformly distributed in the feasible space (Lahdelma and Salminen, 2006). As noted in Kadziński et al. (2017), it is in line with the spirit of robustness analysis, where each feasible weight vector is equally authorized to make some outcome nonnecessary or possible, or shift the extreme bounds.

The distribution of different efficiency results can be captured with the stochastic acceptability indices quantifying the shares of feasible weights confirming a given outcome. We consider the following four types of indices:

- *Distance acceptability interval index (DAII)* (DMU_o, b_i) is the share of feasible weights for which the distance (to the efficiency frontier) of DMU_o to the best unit belongs to the interval $b_i = (b_{i,*}, b_i^*]$, being one of the B buckets partitioning the range $[0,1]$ so that $\bigcup_{i=1}^B b_i = [0, 1]$, $b_i \cup b_j = \emptyset$, $i \neq j$, and b_1 is left-closed, that is, $b_1 = [b_{1,*} = 0, b_1^*]$ (by default, we assume that $b_i^* - b_{i,*} = b_{i+1}^* - b_{i+1,*}$, $i = 1, \dots, B - 1$).
- *Efficiency acceptability interval index (EAII)* (DMU_o, b_i) is the share of feasible weights for which the efficiency (in terms of comprehensive score) of DMU_o , E_o , belongs to the interval b_i .
- *Efficiency rank acceptability index (ERAI)* (DMU_o, r) is the share of feasible weights for which DMU_o attains r th rank (in terms of comprehensive score).
- *Pairwise efficiency outranking index (PEOI)* (DMU_o, DMU_k) is the share of feasible weights for which DMU_o attains at least as good efficiency as DMU_k ($E_o \geq E_k$) (in terms of comprehensive score).

Also, by averaging the measures observed for all feasible weight vectors derived with the Monte Carlo simulations, we may estimate for DMU_o its expected distance $Ed(DMU_o)$ to the efficient DMU, expected efficiency $EE(DMU_o)$, and expected rank $ER(DMU_o)$. These measures can be used to impose a complete order on the considered set of DMUs (Labijak-Kowalska and Kadziński, 2021). Their analysis is beneficial in decision problems with modest stakes or relatively

rich weight constraints, when the average performance or expected values may be used for deriving a decision recommendation representative for the entire set feasible weights.

In Section A1, we illustrate how such stochastic acceptability indices and expected efficiency measures are computed for the study considered in Section 3. To keep this illustration concise, we use a limited set of 10 samples. On the contrary, the results reported for the case study in the main paper are derived from the analysis of 10,000 uniformly distributed weight vectors.

The proposed robustness analysis based on mathematical programming and the Monte Carlo simulations have been made available on the open-source software platform *diviz* (Meyer and Bigaret, 2012). Each method was implemented as an independent module. These modules accept inputs and provide results in the XMCD standard, enabling combining them into complex workflows and visualizing the results using other modules available on *diviz*.

2.5. Selection of a common vector of weights based on the outcomes of robustness analysis

In the traditional DEA models, for each DMU, we select a potentially different weight vector that reflects the most advantageous performance scenario for this unit. While this way of proceeding is useful for verifying the efficiency status of different DMUs, it may prevent a justifiable ranking or a selection of the best units due to the lack of a common base for their comparison (Contreras, 2020). In turn, robustness analysis is oriented toward summarizing the results of comparing the DMUs on all feasible input and output weights, hence offering multiple, possibly infinitely many, bases for joint consideration of all units. Even though such results are useful for understanding the stability of results, some users may find them challenging to understand, mainly due to the multiplicity of weight vectors that serve as the basis for conducting the robustness analysis.

In some applications, it might be more appropriate to consider the same basis for evaluating the DMUs, namely by selecting a common vector of weights for evaluating all DMUs. In this way, all units can be ranked on a unified scale, which increases the discrimination power compared to the classical DEA models. The idea of selecting a common vector of weights was introduced by Charnes et al. (1989), quickly finding its first applications in the evaluation of highway maintenance patrols (Cook et al., 1990) and farms in Kansas (Thompson et al., 1990). Over the last decades, multiple methods for determining a common vector of weights have been proposed. These approaches build on the concepts of ideal and anti-ideal alternatives, weighting schemes, cross-efficiency analysis, incorporating the DM's preferences, evaluating only a proper subset of DMUs, statistical analysis, or game theory (Contreras, 2020).

This section introduces the novel procedures for selecting a common vector of weights based on the analysis of results derived with robustness analysis. Overall, we aim at selecting a single weight vector representing the whole set of feasible input and output weights. Our purpose is to find a vector that matches as well as possible the results deemed to be robust. In particular, if the robust results warrant concluding that some DMU_o is better than some DMU_k , then the difference between the efficiency scores of these two DMUs should be enhanced. This will depend on the truth of a specific robust relation (let us denote it by \succ^W), confirming the evident advantage of one DMU over another given the results attained for all feasible weights. On the other hand, we can point out the pairs of DMUs for which the efficiency difference should be small due to the ambiguity in their comparison, given all input and output weights. Such pairs are incomparable (R^W) in terms of the

Table 1

Conditions justifying the truth of the robust preference \succ^W and incomparability R^W relations

Result	$DMU_o \succ^W DMU_k$	$DMU_l R^W DMU_p$
\succ_E^N	$DMU_o \succ_E^N DMU_k$ and not $(DMU_k \succ_E^N DMU_o)$	not $(DMU_l \succ_E^N DMU_p)$ and not $(DMU_p \succ_E^N DMU_l)$
EE	$EE(DMU_o) - EE(DMU_k) > t_{EE}$	$ EE(DMU_o) - EE(DMU_k) \leq t_{EE}$
ER	$ER(DMU_o) - ER(DMU_k) > t_{ER}$	$ ER(DMU_o) - ER(DMU_k) \leq t_{ER}$
$PEOI$	$PEOI(DMU_o, DMU_k) - PEOI(DMU_k, DMU_o) > t_{PEOI}$	$ PEOI(DMU_l, DMU_p) - PEOI(DMU_p, DMU_l) \leq t_{PEOI}$

robust relation \succ^W . Thus interpreted, the selected common vector of weights is representative for all feasible weight vectors in the sense of the robustness concern.

The outcomes discussed in Sections 2.2 and 2.4 provide diverse bases for defining the conditions underlying the truth or falsity of the robust relation \succ^W . In this paper, we will refer to four possibilities that build on the necessary preference relation (\succ_E^N), expected efficiency scores (EE s) and ranks (ER s), and $PEOIs$. The respective conditions needed for establishing relations \succ^W and R^W are defined in Table 1. For example, when referring to \succ_E^N , one unit can be judged as univocally more advantageous than another if it is necessarily preferred to it, confirming that its efficiency is at least as good for all feasible weights. On the contrary, the comparison based on \succ_E^N can be judged ambiguous if a given pair of units is incomparable in terms of \succ_E^N . This means that for at least one feasible weight vector, one unit is judged more efficient, whereas, for some other input and output weights, the relation is inverse. Furthermore, when referring to the EE s and ER s, we can judge one unit as stochastically preferred to another if its expected efficiency or rank is better by some pre-defined threshold, t_{EE} or t_{ER} , specifying the minimal difference in expected results justifying an evident advantage. When such a threshold is not exceeded, we may assume that the difference is negligible. Finally, as far as $PEOIs$ are concerned, the truth of a robust preference relation \succ^W is well motivated when the share of feasible weights for which DMU_o is more efficient than DMU_k is greater than the share of weights for which the relation is inverse by more than threshold t_{PEOI} . By default, thresholds t_{EE} , t_{ER} , and t_{PEOI} are set to zero. However, the user can also set them to some positive values, hence imposing more demanding requirements for instantiating \succ^W as well as a greater tolerance for establishing R^W .

The selection of a common vector of weights is conducted by attaining the two targets lexicographically. First, we maximize the minimal difference between efficiency scores for pairs of units related by \succ^W , that is,

$$\text{Maximize } \alpha \tag{9}$$

s.t.

$$\left. \begin{array}{l} \text{for } (DMU_o, DMU_k) \in \mathcal{D} \times \mathcal{D} : DMU_o \succ^W DMU_k : \\ \sum_{q=1}^Q w_q u_q(DMU_o) - \sum_{q=1}^Q w_q u_q(DMU_k) \geq \alpha, \\ \mathcal{W}. \end{array} \right\}$$

Let us denote the optimal solution of the above LP problem by α^* . Second, we minimize the maximal difference between efficiency scores for pairs of units related by $R^{\mathcal{W}}$, that is,

$$\text{Minimize } \beta \tag{10}$$

s.t.

$$\left. \begin{array}{l} \text{for } (\text{DMU}_l, \text{DMU}_p) \in \mathcal{D} \times \mathcal{D} : \text{DMU}_l R^{\mathcal{W}} \text{DMU}_p : \\ \sum_{q=1}^Q w_q u_q(\text{DMU}_l) - \sum_{q=1}^Q w_q u_q(\text{DMU}_p) \leq \beta, \\ \sum_{q=1}^Q w_q u_q(\text{DMU}_p) - \sum_{q=1}^Q w_q u_q(\text{DMU}_l) \leq \beta, \\ \text{for } (\text{DMU}_o, \text{DMU}_k) \in \mathcal{D} \times \mathcal{D} : \text{DMU}_o \succ^{\mathcal{W}} \text{DMU}_k : \\ \sum_{q=1}^Q w_q u_q(\text{DMU}_o) - \sum_{q=1}^Q w_q u_q(\text{DMU}_k) \geq \alpha^*, \\ \mathcal{W}. \end{array} \right\}$$

The common vector of weights selected in this way can be used to order all units from the best to the worst. The obtained ranking emphasizes the outcomes following the use of all feasible weights, which contributed to the selection of an underlying representative vector of weights. It positively affects the accuracy of the provided results while extending the robustness analysis in the capacity to explain its outcomes. The user can analyze the computed weights and efficiency scores, which is more understandable than examining the necessary, extreme, expected, or stochastic outcomes. Note that this idea has not yet been explored in the context of DEA, even though it has been successfully applied in MCDA (Kadziński et al., 2012a).

3. Case study: efficiency evaluation of emergency department physicians

In this section, we discuss the application of the proposed method for evaluating the performance of 20 full-time ED physicians. Data used in the study came from a sufficiently long period of time (12 months) and was controlled for a case-mix. A detailed description of the case study setting can be found in Fiallos et al. (2017).

We consider the following three inputs, reflecting the essential resources consumed by the physicians in the process of managing patients in the ED:

- i_1 —an average encounter time per patient visit (AVG_MDTIME_PAT), which is defined as an average number of minutes between the first contact of the physician with the patient and the moment a disposition decision is made and recorded on a patient's chart;
- i_2 —an average number of laboratory tests per patient visit (AVG_LAB_PAT) when diagnosing a patient;

Table 2

Input and output values for the 20 physicians given complaint group *G1* (abdominal pain and constipation) (Fiallos et al., 2017)

MD	i_1 —AVG_MDTIME_PAT	i_2 —AVG_LAB_PAT	i_3 —AVG_RAD_PAT	o_1 —RATE_NR72
MD1	2.026	2.760	0.920	1.000
MD2	1.959	2.381	0.774	0.961
MD3	2.223	2.333	0.643	0.905
MD4	1.884	1.823	0.661	0.952
MD5	1.511	0.857	0.487	0.952
MD6	1.456	1.330	0.648	0.978
MD7	1.903	1.877	0.596	0.956
MD8	1.704	1.730	0.678	0.939
MD9	1.708	1.927	0.657	0.968
MD10	1.979	1.508	0.820	0.922
MD11	1.652	1.618	0.592	0.981
MD12	2.169	1.863	0.608	0.961
MD13	1.634	1.538	0.786	0.979
MD14	1.745	2.117	0.738	0.942
MD15	1.594	1.548	0.602	0.957
MD16	2.311	1.538	0.462	0.974
MD17	1.962	1.748	0.557	0.948
MD18	1.804	1.590	0.723	0.977
MD19	1.567	1.487	0.601	0.937
MD20	1.435	1.198	0.568	0.969

- i_3 —an average number of radiology orders per patient visit (AVG_RAD_PAT) used in the diagnosis.

Indeed, one can expect that an efficiently working physician arrives at the correct diagnosis in a shorter time and ordering fewer laboratory tests and radiology orders than a less efficient one. As an output (o_1), we will consider each physician's quality of care measured by the rate of nonreturn patient visits within 72 hours of discharge (RATE_NR72). Such a value has been traditionally considered one of the most informative indicators of the physicians' performance (Hung and Chalut, 2008).

Patients have a variety of reasons for visiting an ED. Given different complaint groups, one may observe variations in the clinical practices and different levels of the available resources such as time, tests, or orders. For this reason, the efficiency of physicians should be evaluated individually for each complaint type, representing a different clinical and diagnostic category. In this case study, our primary focus is on a group of patients complaining (*G1*) about abdominal pain and constipation. The input and output values for this group are presented in Table 2. In this context, we will discuss the results of robustness analysis obtained with mathematical programming, the Monte Carlo simulation, and common sets of weights selected using different procedures. We will also consider two other complaint groups—fever (*G2*) and lower or upper extremity injury, head injury, and laceration/puncture (*G3*). The three groups will serve as the basis for the multiscenario robustness analysis. The descriptive statistics of inputs and outputs for all considered groups are

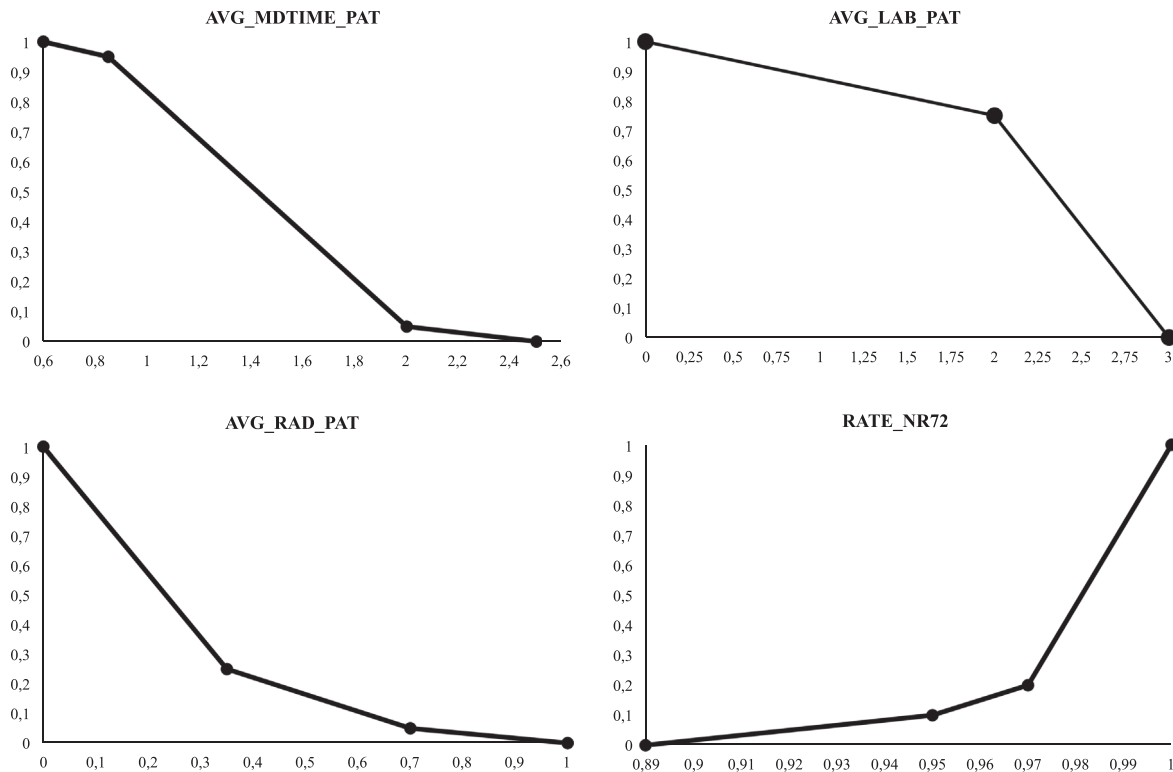


Fig. 1. Marginal value functions for the inputs and output used to evaluate the performance of ED physicians (x-axis—performances; y-axis—marginal value).

given in Section A2. To keep the main paper concise, some other data and results are also presented or discussed in the Appendices.

The marginal functions that will be used in the value-based efficiency analysis are presented in Fig. 1. They have been elicited from an independent medical expert using a direct questioning technique. He took into account performance ranges for each factor, the per-factor preferences, and the performances' distribution. This led to defining the convex functions for i_3 and o_1 , a concave function for i_2 , and a sigmoid-like function for i_1 . Moreover, to prevent the dominating role of any factor on the final results, their weights have been constrained to at most 0.5 (i.e., $w_q \leq 0.5$, $q \in \{i_1, i_2, i_3, o_1\}$). We incorporated the latter assumption to avoid scenarios in which a physician is deemed efficient simply because of excelling at only one aspect of the clinical role while being ineffective at all other aspects.

Note that in the original case study, physicians' performance was analyzed using a more traditional SBM-SWAT VRS model that considers a single most advantageous weight vector for each DMU (Fiallos et al., 2017). The results presented in Fiallos et al. (2017) take the form of precise efficiency scores for each physician and each complaint group, providing somewhat limited and straightforward insights. In the following subsections, we discuss the insights derived from the analysis of all feasible weights offering means for identifying overall good or bad performers and

applying individual common sets of weights forming the basis for deriving univocal and well-justified ranking of physicians. In this perspective, we increase the discriminative power of efficiency results compared to the traditional methods and address the criticism leveled against the way the efficiency scores are computed in these approaches when analyzing only the input/output weights, which are the most favorable to each DMU. Moreover, we focus on different perspectives on the efficiency of physicians while acknowledging that ranks and pairwise preference relations are more interpretable to nonspecialists in DEA. When it comes to multiple scenarios, we present the aggregated results summarizing physicians' performance for different complaint groups instead of simply displaying the numerical outcomes for each considered scenario individually.

3.1. Robust efficiency results for complaint group G_1 : abdominal pain and constipation

This section presents the robust results for complaint group G_1 concerning abdominal pain and constipation. First, we discuss robustness analysis outcomes referring to the efficiency scores (the discussion on the rank-related perspective and pairwise preference relations is provided in the Appendices). Second, we present the common sets of weights and the underlying rankings of physicians.

3.1.1. Distances to the efficient unit and efficiency scores

This section discusses the robustness of distances to the efficient physician and efficiency scores for the set of 20 physicians (further referred to as MD1, etc.). In Table 3, we present the extreme distances (columns d_* and d^*) and scores (columns E_* and E^*). The minimal distance d_* is equal to 0 for six physicians: MD1, MD5, MD6, MD11, MD16, and MD20. They are deemed efficient because they attain the greatest efficient score for at least one feasible weight vector. On the other extreme, even for the best scenario for MD2 and MD3, their minimal distances to the efficient physician are quite large (0.1836 and 0.1911, respectively). This implies that they are far from working efficiently.

The maximal efficiency score E^* is strongly correlated with the minimal distance d_* . This is understandable because if some physician acts efficiently or (s)he is close to being efficient, this should be due to attaining a relatively high efficiency score in the most favorable scenario. The greatest efficiency scores are attained by MD20 (0.6712) and MD6 (0.6547). It is worth noting that E^* for the efficient physician MD1 (0.5900) is lesser than E^* for the inefficient physicians: MD13 (0.6239), MD18 (0.5940), and MD19 (0.6015). This confirms the importance of analyzing the relative distances rather than absolute scores when deciding about efficiency.

When considering the least favorable scenarios, the best maximal distances d^* are between the two efficient physicians, MD11 (0.2041) and MD6 (0.2601), and inefficient MD13 (0.2688). This indicates that even in the most pessimistic scenarios for these physicians, the differences in their efficiencies are relatively small in terms of their scores on a scale of comprehensive value. The worst maximal distances d^* are for MD3 (0.4974) and MD1 (0.5575), being about twice as large as for the best performing physicians. When it comes to the minimal efficiency scores E_* , the best physicians are also MD6 (0.2465) and MD11 (0.2170). In turn, the least performing ones are MD1 (0.0304) and MD3 (0.0264), with efficiency scores very close to zero in the most pessimistic scenario.

Table 3

Extreme and expected values of distances to the efficient unit and efficiency scores for all considered physicians

MD	d_*	d^*	Ed	E_*	E^*	EE
1	0.0000	0.5575	0.1601	0.0304	0.5900	0.3169
2	0.1836	0.4103	0.2844	0.0599	0.3096	0.1882
3	0.1911	0.4974	0.3096	0.0264	0.2914	0.1619
4	0.0921	0.4155	0.1945	0.0911	0.4565	0.2785
5	0.0000	0.3658	0.0675	0.1409	0.6628	0.4048
6	0.0000	0.2601	0.0196	0.2465	0.6547	0.4552
7	0.0764	0.3957	0.1854	0.1177	0.4477	0.2873
8	0.0950	0.4345	0.1667	0.0721	0.5327	0.3061
9	0.0812	0.3744	0.1436	0.1323	0.5188	0.3297
10	0.0967	0.4650	0.2275	0.0417	0.4390	0.2457
11	0.0000	0.2041	0.0380	0.2170	0.6455	0.4370
12	0.0755	0.4297	0.2031	0.0678	0.4611	0.2699
13	0.0153	0.2688	0.0646	0.1861	0.6239	0.4108
14	0.1377	0.4415	0.2089	0.0652	0.4559	0.2638
15	0.0543	0.3862	0.1157	0.1205	0.5871	0.3572
16	0.0000	0.3609	0.1373	0.1025	0.5572	0.3361
17	0.0634	0.4355	0.1953	0.0882	0.4566	0.2772
18	0.0382	0.2903	0.1094	0.1248	0.5940	0.3656
19	0.0563	0.4142	0.1226	0.0925	0.6015	0.3499
20	0.0000	0.3465	0.0543	0.1602	0.6712	0.4188

To judge the stability of efficiency results for all feasible weights, we can refer to the distance and efficiency intervals' widths. On the one hand, the difference between d^* and d_* is the smallest for MD11 (0.2041), confirming the robustness of its relatively high-performance evaluation. On the other hand, for MD1, this difference is the greatest (0.5575), indicating high dependence of results on the selected input and output weights.

To expand the analysis of extreme distances and efficiency scores, we will estimate their distributions using Monte Carlo simulation (see Section A3 for the detailed results), considering 10 equally distributed buckets, from [0.0, 0.1] to (0.9, 1.0]. Note that the methods would work with any other arbitrarily specified subranges. Such distributions are useful for identifying the physicians consuming all their inputs and producing outputs efficiently, independently of the selected factor weights, or those physicians who are more oriented toward optimizing an individual input or output. Let us emphasize that smaller values are better when considering the distances, and larger values are better when considering the efficiency scores.

The distance of MD6 and MD11 to the efficient physician is lower than 0.1 for more than 95% weight vectors. This confirms that these physicians perform efficiently or are very close to being efficient for the vast majority of scenarios. Furthermore, even though MD16 is efficient, its distance from the efficient physician is most often between 0.1 and 0.2 (51.2%), and only for 29.5% weights, it lies in the interval [0.0, 0.1]. This suggests that MD16 cannot optimize all inputs and outputs equally well. The analysis of $DAIIs$ and $EAIIs$ is also helpful to identify the underperforming physicians. For example, the efficiency scores for MD2 and MD3 are at most 0.2 for, respectively, 57.3% and 68.4% feasible weight vectors, hence confirming their low performance in terms of the efficiency of provided care.

Table 4
Common sets of weights selected using four different procedures

Procedure	w_{i_1}	w_{i_2}	w_{i_3}	w_{o_1}
\succsim_E^N	0.46541	0.21630	0.11753	0.20076
<i>ER</i>	0.23715	0.24646	0.26183	0.25456
<i>EE</i>	0.36510	0.30947	0.00000	0.32543
<i>PEOI</i>	0.25502	0.19246	0.29040	0.26213

To construct a complete ranking of physicians without using a common vector of weights, we can use the expected distances to the efficient unit (*Ed*) and expected efficiencies (*EE*). These metrics are summarized in Table 3. They impose the same orders on the set of physicians under consideration. On the one hand, the top-ranked physicians are MD6 (*Ed* = 0.0196 and *EE* = 0.4552) and MD11 (*Ed* = 0.038 and *EE* = 0.437). For them, the difference to the best physician is, on average, very low, which confirms their position as overall good performers. On the other hand, the bottom-ranked physicians are MD2 (*Ed* = 0.2844 and *EE* = 0.1882) and MD3 (*Ed* = 0.3096 and *EE* = 0.1619), characterized by larger expected distances to the best physicians and lower expected efficiencies.

In general, the analysis of extreme distances and efficiency scores allows distinguishing the MDs exhibiting universal good practices to follow. These include units that attain favorable results for the wide spectrum of feasible weights. In this perspective, MD6 and MD11 can be considered for others as the benchmarks. Other MDs that are efficient only under specific conditions can be judged more niche (see, e.g., MD1 and MD16). These results are also helpful in discriminating between the inefficient DMUs. On the one hand, MDs with favorable extreme distances and scores have the most significant potential for becoming efficient. Therefore, the management may implement the corrective plan for units such as MD13 and MD18 in the first order. On the other hand, high distances and low scores indicate the MDs for which becoming efficient would be the most challenging, and the corrective actions need to be distributed over a longer-term (see, e.g., MD2 and MD3).

An analogous discussion on the robustness of efficiency ranks and pairwise preference relations is presented in Sections A4 and A5.

3.1.2. Analysis of rankings obtained by applying the common sets of weights

This section reports the results obtained using four procedures for selecting the common vector of weights presented in Section 2.5. They build on the expected efficiencies *EEs* (see Section 3.1.1), expected ranks *ERs* (see Section A4), the necessary preference relation \succsim_E^N , or *PEOIs* (see Section A5). We parameterize the procedures with the following thresholds justifying the truth of a robust preference relation: $t_{ER} = 0.5$, $t_{EE} = 0.1$, and $t_{PEOI} = 0.15$. Hence, to justify an evident advantage in performance of one physician over another, his/her expected rank should be better by more than 0.5, or the expected efficiency should be greater by more than 0.1, or the share of feasible input/output weights confirming better performance should be greater by more than 15% than the share of weights confirming worse performance.

In Table 4, we present the common sets of weights selected using four different procedures. For example, when considering the weights chosen based on the analysis of \succsim_E^N , the highest priority

Table 5

Efficiency scores and ranks attained by physicians for the common sets of weights selected using four different procedures

Procedure	$\underset{E}{\succsim}^N$		<i>ER</i>		<i>EE</i>		<i>PEOI</i>	
	Efficiency	Rank	Efficiency	Rank	Efficiency	Rank	Efficiency	Rank
1	0.2633	13	0.3137	11	0.3984	10	0.3127	8
2	0.1742	19	0.1832	19	0.2241	19	0.1618	19
3	0.1358	20	0.1578	20	0.1731	20	0.1339	20
4	0.2631	15	0.2706	14	0.3261	13	0.2343	14
5	0.4368	3	0.3956	5	0.4701	5	0.3609	5
6	0.4941	1	0.4444	1	0.5662	1	0.4133	1
7	0.2631	14	0.2802	13	0.3251	14	0.2453	13
8	0.3244	10	0.2971	12	0.3720	11	0.2622	12
9	0.3407	9	0.3210	10	0.3984	9	0.2886	11
10	0.2207	18	0.2372	18	0.2927	18	0.1958	18
11	0.4347	4	0.4279	2	0.5251	2	0.3975	2
12	0.2245	17	0.2632	16	0.2999	17	0.2265	17
13	0.4238	5	0.4002	4	0.5160	3	0.3669	4
14	0.2819	11	0.2559	17	0.3243	15	0.2265	16
15	0.3852	7	0.3481	7	0.4278	7	0.3152	7
16	0.2669	12	0.3303	9	0.3567	12	0.2947	10
17	0.2410	16	0.2706	15	0.3024	16	0.2343	15
18	0.3510	8	0.3562	6	0.4481	6	0.3208	6
19	0.3853	6	0.3407	8	0.4194	8	0.3073	9
20	0.4669	2	0.4088	3	0.5063	4	0.3767	3

is assigned to i_1 , whereas the lowest priority is attributed to i_3 . On the contrary, the values of weights selected based on *ERs* are more balanced, ranging between 0.23715 (for i_1) and 0.26183 (for i_3).

The respective efficiency scores and ranks for the 20 physicians are given in Table 5. These scores are derived from the lexicographic optimization of two targets—maximization of the efficiency difference for pairs of physicians related by the robust preference relations and minimization of such a difference for pair incomparable in terms of this relation.

Let us discuss in detail the results built on *ER* and $\underset{E}{\succsim}^N$. When it comes to the expected ranks (see Section A4), the three best performing physicians are MD6 (1.860), MD11 (2.914), and MD20 (3.682), whereas the three bottom-ranked physicians are MD10 (17.347), MD2 (18.669), and MD3 (19.640). The expected ranks' analysis is the basis for selecting a common vector of weights. For example, MD6 should be preferred to MD20, which, in turn, should be judged better than MD5, etc., according to the common weight vector to be chosen. Solving the LP problem (Section 2.5), the minimal efficiency difference for pairs with expected ranks differing by more than 0.5 is positive (0.00738). This means that the derived rankings reflect the order of physicians implied by *ERs*. For example, MD6 is ranked first with an efficiency of 0.444, and MD3 is ranked last with an efficiency of 0.1578. Thus, the expected results derived from the analysis of all feasible weights have been captured with a single common weight vector: [0.23715, 0.24646, 0.26183, 0.25456]. Moreover, the derived ranking can be seen as a synthetic representation of the expected results derived from the stochastic analysis.

Table 6
Values of Kendall's τ coefficient for all pairs of rankings obtained with different procedures

Procedure	$\tilde{\gamma}_E^N$	ER	EE	$PEOI$
$\tilde{\gamma}_E^N$	1.000	0.842	0.853	0.821
ER	0.842	1.000	0.926	0.958
EE	0.853	0.926	1.000	0.926
$PEOI$	0.821	0.958	0.926	1.000

In the same spirit, the efficiency scores built on $\tilde{\gamma}_E^N$ (see Section A5) allowed flattening, in a reasonable way, the graph of the necessary preference relation determined with mathematical programming. The minimal efficiency difference for pairs related by $\tilde{\gamma}_E^N$ is 0.04243. Hence, the procedure succeeded in reflecting the preference confirmed by all feasible weights in a complete order imposed by applying a single weight vector: [0.46541, 0.21630, 0.11753, 0.20076]. For example, such an advantage can be observed for the following pairs: (MD6, MD15), (MD15, MD8), (MD8, MD14), and (MD14, MD3). Furthermore, the physicians who are necessarily preferred to many other physicians attain the best scores and ranks according (see MD6 (1), MD20 (2), MD5 (3), MD11 (4), and MD13 (5)). On the contrary, the physicians necessarily outperformed by many others are ranked at the bottom (see MD10 (18), MD2 (19), and MD3 (20)). Interestingly, MD1, being incomparable in terms of $\tilde{\gamma}_E^N$ with any other physician, is ranked 13th, hence attaining an intermediate position. Overall, the analysis of such a ranking supports the comprehension of the necessary preference relation, making comparisons among the physicians more clear and the entire order well justified due to its roots in the outcomes observed for all feasible weight vectors.

The rankings constructed by the four procedures (see Table 5) are very similar. In Table 6, we present the values of Kendall's τ coefficient (Winkler and Hays, 1985) for all pairs of obtained rankings. For example, MD6 and MD3 are ranked at, respectively, the very top and very bottom by all procedures. The slight differences between the ranking produced by different procedures are the result of different tolerance levels that were used. On the one hand, the necessary preference relations left many pairs of physicians incomparable, whereas the expected ranks coupled with $t_{ER} = 0.5$ allowed comparing almost all pairs of physicians. On the other hand, requiring that physician's expected efficiency is better than another by more than 0.1 is clearly more limiting than requiring the difference in expected ranks to be greater than 0.5. Consequently, different numbers of pairs of physicians were considered in the two phases of lexicographic optimization. While this had an impact on the rankings, the subsets of the best, medium, and the worst performers stay the same.

The discussed rankings are also strongly correlated with the one presented in Fiallos et al. (2017), derived using the SBM-SWAT VRS model. The correlation coefficients range from 0.611 to 0.723 when considering the ranking based on EE or $\tilde{\gamma}_E^N$, respectively. When comparing the four rankings with the order reported in Fiallos et al. (2017), the positions attained by MD2, MD6, MD10, MD12, MD15, and MD20 differ by at most 2. The greatest differences are observed for MD1, MD3, MD11, and MD16 (up to 10, 6, 7, and 6 positions, respectively). The reasons underlying these differences have various origins. For example, we demonstrated that the performance of MD1 highly depends on the selected weight vector, while it was ranked at the bottom in Fiallos et al. (2017). Moreover, MD3 was judged as the worst performing physician according to all ranking

Table 7

The possible-necessary and possible-possible intervals of distances to the efficient physician, efficiency scores, and ranks based on the analysis of three complaint groups

MD	$[d_{*,a,S}^P, d_{a,S}^{*,P}]$	$[d_{*,a,S}^N, d_{a,S}^{*,N}]$	$[E_{*,a,S}^P, E_{a,S}^{*,P}]$	$[E_{*,a,S}^N, E_{a,S}^{*,N}]$	$[R_{*,a,S}^P, R_{a,S}^{*,P}]$	$[R_{*,a,S}^N, R_{a,S}^{*,N}]$
1	[0.0000, 0.5575]	[0.0631, 0.3784]	[0.0304, 0.9623]	[0.3095, 0.5900]	[1, 20]	[12, 18]
2	[0.0861, 0.4292]	[0.1836, 0.3437]	[0.0599, 0.7936]	[0.2861, 0.3096]	[13, 20]	[14, 20]
3	[0.0607, 0.4974]	[0.1911, 0.4653]	[0.0264, 0.8663]	[0.1735, 0.2914]	[9, 20]	[18, 20]
4	[0.0072, 0.4938]	[0.1277, 0.1317]	[0.0911, 0.9928]	[0.3850, 0.4565]	[2, 20]	[12, 18]
5	[0.0000, 0.4423]	[0.0839, 0.3174]	[0.1409, 0.8954]	[0.2710, 0.6628]	[1, 17]	[7, 8]
6	[0.0000, 0.4358]	[0.0000, 0.1670]	[0.2465, 0.9486]	[0.4215, 0.6547]	[1, 14]	[1, 5]
7	[0.0444, 0.4231]	[0.1057, 0.1830]	[0.1177, 0.8876]	[0.4054, 0.4477]	[6, 18]	[11, 12]
8	[0.0019, 0.4345]	[0.0950, 0.3191]	[0.0721, 0.9359]	[0.3579, 0.5327]	[2, 17]	[8, 10]
9	[0.0461, 0.4367]	[0.0812, 0.2946]	[0.1323, 0.9001]	[0.2938, 0.5188]	[5, 18]	[9, 15]
10	[0.0172, 0.4650]	[0.0967, 0.2591]	[0.0417, 0.8471]	[0.3318, 0.4390]	[3, 20]	∅
11	[0.0000, 0.4421]	[0.0706, 0.2041]	[0.2170, 0.8971]	[0.3515, 0.6455]	[1, 20]	[6, 8]
12	[0.0000, 0.4445]	[0.0755, 0.1401]	[0.0678, 0.9340]	[0.4483, 0.4611]	[1, 18]	∅
13	[0.0153, 0.3145]	[0.0257, 0.1597]	[0.1861, 0.9150]	[0.4287, 0.6239]	[2, 14]	[3, 7]
14	[0.0507, 0.4415]	[0.1377, 0.1791]	[0.0652, 0.8690]	[0.4106, 0.4559]	[4, 19]	[10, 16]
15	[0.0000, 0.3862]	[0.0543, 0.3312]	[0.1205, 0.9686]	[0.3120, 0.5871]	[1, 18]	[5, 11]
16	[0.0000, 0.4585]	∅	[0.1025, 1.0000]	[0.5094, 0.5572]	[1, 20]	[2, 6]
17	[0.0602, 0.4355]	[0.0910, 0.1881]	[0.0882, 0.8724]	[0.4003, 0.4566]	[3, 18]	[6, 18]
18	[0.0382, 0.4230]	[0.0915, 0.2903]	[0.1248, 0.8061]	[0.3039, 0.5940]	[4, 19]	[12, 14]
19	[0.0269, 0.4142]	[0.0584, 0.1659]	[0.0925, 0.9056]	[0.4225, 0.6015]	[3, 15]	[4, 9]
20	[0.0000, 0.4541]	[0.0139, 0.3465]	[0.1602, 0.9711]	[0.2278, 0.6712]	[1, 18]	[2, 8]

methods considered in this paper. This is implied by its unfavorable evaluation for the vast majority of feasible weights, which follows the transformation of its performances into marginal values using the functions presented in Fig. 1. However, according to Fiallos et al. (2017), five other physicians were judged worse than MD3.

3.2. Multiscenario robustness analysis for different complaint groups

In this section, we present the aggregated results of robustness analysis for the three complaint groups related to abdominal pain and constipation ($G1$), fever ($G2$), and lower or upper extremity injury, head injury, and laceration/puncture ($G3$). The input and output values for groups $G2$ and $G3$ are given in Section A6. The analysis of pairwise-oriented outcomes is provided in Section A7. In the main paper, we focus on the robust intervals of distances to the best physician, efficiencies, and ranks.

To derive the aggregated score- and rank-related results for three complaint groups, we conducted a robustness analysis for each of them individually and introduced a second level of certainty to capture the stability of outcomes for physicians treating patients from different groups. In Table 7, we present the extreme distances to the efficient physician, efficiency scores, and ranks obtained in that way. These marked as necessary (N) indicate the values obtained for all complaint groups, while the ones denoted as possible (P) specify the values obtained for at least one group.

The lower bound of the possible distance interval $[d_{*,o,S}^P, d_{o,S}^{*,P}]$ is equal to 0 for eight physicians: MD1, MD5, MD6, MD11, MD12, MD15, MD16, and MD20. These physicians perform efficiently, treating at least one complaint group. Moreover, MD6 is the only physician for whom the lower bound of the necessary distance interval $[d_{*,o,S}^N, d_{o,S}^{*,N}]$ is 0. This confirms its efficiency for all three complaint groups. The next two best results are attained by MD20 (0.0139) and MD13 (0.0257), which means that they are nearly efficient for all considered settings. In turn, for MD16, the intersection of the distances to the efficient physician over all groups is empty. Such an outcome indicates that MD16's performance strongly depends on the group. (S)he performed quite well for one group and all feasible input and output weights and poorly for some other group.

The possible-possible intervals of efficiency scores $[E_{*,o,S}^P, E_{o,S}^{*,P}]$ are wide for all physicians. The minimal width is for MD11 (0.6801), whereas the maximal difference between the extreme scores for different complaint groups is equal to 0.9319 (see MD1). When it comes to the width of the possible-necessary efficiency score interval $[E_{*,o,S}^N, E_{o,S}^{*,N}]$, it is minimal for MD12 (0.0128) and maximal for MD20 (0.4434). The physicians with the greatest width of the possible-necessary interval and the least width of the necessary-necessary interval are the most specialized ones, attaining highly variable results for different complaint groups.

Similar conclusion can be derived from the analysis of multiscenario rank intervals (see $[R_{*,o,S}^P, R_{o,S}^{*,P}]$ and $[R_{*,o,S}^N, R_{o,S}^{*,N}]$). The relative performance of MD2 and MD3 is rather poor for all complaint groups. Their best rank for any group is 13 and 9, respectively. For other physicians, the possible-possible rank intervals are rather wide, again confirming their varied performance. In particular, there are three physicians (MD1, MD11, and MD16) who attributed all ranks when considering the three complaint groups.

When considering the possible-necessary rank intervals $[R_{*,o,S}^N, R_{o,S}^{*,N}]$, we can observe that for MD10 and MD12, there is no single rank attained for all complaint groups. The best results are observed for MD5 who attained ranks in the interval $[1, 5]$ for all scenarios. Similarly, in the most favorable scenario, MD16 and MD20 are ranked at least second ($R_{*,o,S}^N = 2$) for all complaint groups. On the contrary, MD1 is ranked only 12th for one group ($R_{*,o,S}^N = 12$). Given its efficiency for some other group ($R_{*,o,S}^P = 1$), this means that the performance of MD1 mostly depends on the selected priorities and evaluation scenario.

In Section A8, we summarize the results derived for each physician with the proposed robustness analysis framework for a single scenario referring to complaint group $G1$ and multiple scenarios concerning groups $G1$, $G2$, and $G3$. Specifically, we refer to the ranks attained by each physician according to various measures.

4. Conclusions and implications

We presented a novel robustness analysis framework for DEA incorporating a value-based additive efficiency model. The basic framework incorporates mathematical programming techniques and the Monte Carlo simulation to exploit all feasible input and output weights. These methods derive two types of results concerning four perspectives relevant to the analysis. One type of results, extreme outcomes, captures exact outcomes observed for the most and the least advantageous weight vectors for a given DMU or instantiated for all or at least one feasible weight vector. Another type

of results, stochastic acceptability indices, quantify the share of feasible weight vectors supporting some conclusions. The four accounted perspectives concern efficiency scores, distances from the efficient unit, ranks, and pairwise efficiency preference relations. Such outcomes provide rich information on the stability of efficiency outcomes from the complementary perspectives that focus on the DMUs assessed individually, compared pairwise, or collated with all remaining units in the analyzed set. To facilitate the application of these methods in practice, we created an open-source system implementing them on the *diviz* platform.

In addition, the primary framework was extended in two ways. On the one hand, we introduced the procedures for selecting the common vector of weights. These procedures incorporate robustness by exploiting stability analysis outcomes to define the score differences that should be emphasized in the ranking constructed with the chosen weight vector. One may either maximize the differences between efficiencies for pairs of DMUs for which an evident advantage of either of them can be observed given results attained for all feasible weight vectors or minimize such a difference if the results of such a comparison are not univocal. Specifically, we discussed the procedures exploiting the necessary efficiency preference relation, expected efficiencies, expected ranks, or *PEOIs*. On the other hand, we adjusted the robustness analysis framework to a multiscenario setting, in which the same DMUs are evaluated under different conditions or from various perspectives. The main innovation consisted of accounting for the second level of certainty, referring to the necessity or possibility of some robust conclusion given multiple relevant scenarios.

The proposed approach was applied to evaluating the performance of the ED physicians, assuming time, laboratory tests, and radiology orders as inputs, and rate of nonreturn visits to the ED within 72 hours as a single output that is a proxy for physicians' performance and the quality of the provided care. The robust results provide multiple implications for both individual physicians and hospital managers. Let us emphasize that due to the specificity of DEA, these conclusions are limited by considering a specific setup involving a particular group of analyzed peers, factors selected as relevant for the analysis, and an adopted efficiency model. Thus, they do not refer to any external standards.

First, the wide intervals of efficiency scores, distances to the efficient physician, and ranks, observed for most physicians for a single complaint group, serve as the evidence for the strong dependence of the physicians' performances on the selected weight vector (i.e., priorities assigned to different inputs and outputs). Such a high variability of results should make the analysts careful with some definitive judgments about the physicians' performance and might help identify the outliers. This variability also puts into question the results obtained with traditional DEA methods taking into account only the most advantageous scenario for each DMU, MCDA approaches, or composite indicators, due to their reliance on a single, often user-defined subjective weight vector or a limited subset of weight vectors.

Second, even though one should not draw strict conclusions about individual physicians' efficiency, the robust results serve as a good starting point for an in-depth investigation. In particular, these outcomes can be used to identify physicians who are markedly better in providing care to a given complaint group. These best performers should have low distances to the efficient physician, high efficiency scores or ranks for most feasible weight vectors, and not be outperformed by one other physician in terms of the necessary preference relation. The physicians satisfying these conditions may be considered a benchmark or "role models." A detailed analysis of their performances can facilitate developing an improvement plan and guidelines for the underperforming ones.

Third, to facilitate communicating the performance assessment results, we provide means for ranking the physicians. On the one hand, such ranking can be determined based on the expected efficiency scores or ranks. They offer an overview of the physicians' average performance (considering different weights), pointing out the overall good performers, niche performers, and lower-performing physicians. On the other hand, the rankings can be determined using a common vector of weights selected to represent the robust results attained for all feasible weight vectors. Such representative weights can also be interpreted as the priorities assigned to different inputs and outputs. They can be used in a practice-oriented model for a given complaint group.

Fourth, the results of robustness analysis are helpful in designing the corrective plans for underperforming physicians. In particular, the necessary preference relation can serve to construct the improvement paths based on the performance of other physicians, who outperform others. Hence, these outcomes may find application in a stepwise benchmarking process. Moreover, when referring to the robust results, the management may formulate detailed and diverse performance targets (e.g., improving inputs and outputs warranting a possible rank in the top three or the necessary preference over some other unit).

Fifth, the outcomes of multiscenario robustness analysis for different complaint groups are useful from individual physicians' and hospital managers' viewpoints. Specifically, we may identify physicians performing well given all complaint groups. They may be treated as universal benchmarks. Other physicians who performed well only for some complaint groups while underperforming for others may be considered "specialists," particularly efficient in managing patients of a given type. Overall, we observed a significant variability of results in the three complaint groups, indicating that medical practices and quality of care vary. From the managerial viewpoint, these outcomes help distinguish physicians into subsets treating patients with different complaints, which can positively affect the overall quality of care. They are also useful for identifying the most difficult complaint groups that are characterized by a low number of efficient physicians and a high number of inefficient physicians.

The main purpose of our research was to show the clinical management insights that can be gained from the robust analytical approach. These insights confirmed some hypotheses (e.g., on the differences between physicians in terms of their efficiencies and in the clinical judgments across the groups), supported common beliefs (e.g., that it is barely possible to excel at only all aspects of the clinical role), and provided answers to some performance-oriented questions (e.g., by identifying specialists or overall good performers). However, having such an approach actually applied in CHEO would require the Research Ethics Board approval and consent of the ED physicians, which was beyond the scope of this study.

Our model's main limitations come from the need to specify the marginal value functions and the lack of indicating precise performance improvements on the particular inputs or outputs that allow attaining efficiency. When it comes to the former, in MCDA, there exist some well-established techniques for eliciting such marginal functions. Moreover, such functions help differentiate between performances on a particular factor based on a given problem's specific features, a set of analyzed DMUs, and management preferences. If such a specification is not possible, one can use a default option of linear marginal value functions. As far as the required improvements are concerned, we instead opt for pointing out the peers from whom one should learn and improvement paths indicating the set of benchmarks.

Let us emphasize that when all components of the proposed methodology are employed simultaneously, the number of results to be considered by decision analysts can be prohibitively large. However, in the context of a real-world application, these components can be limited by accounting for the following three aspects. The first aspect refers to whether the performance of DMUs should be analyzed in single or multiple scenarios (in our paper, these scenarios corresponded to different complaint groups). The second aspect concerns the model exploitation by looking at the robustness of efficiency results or developing a univocal recommendation using a common set of weights emphasizing the robust outcomes. The last aspect refers to a type of output variability (extreme or stochastic) and a perspective on the efficiency analysis (scores, distances, ranks, or preference relations) that should be considered. Having answered such questions, one can limit the scope of the proposed methodological framework to one's own needs.

Several future research directions can be explored. From the application viewpoint, the most interesting one concerns extending the analysis to other complaint groups and more performance measures. In particular, the input- and output-oriented perspectives could be enriched by considering specialist consults and patient satisfaction, respectively. Such data were not available for our study. One could also analyze the impact of a trainee factor on physicians' performance by separately considering the visits when any trainee did not assist them, or junior or senior trainees supported them. From the methodological viewpoint, the proposed robustness analysis framework incorporating a value-based additive efficiency model can be extended to account for the imprecise (interval and ordinal) performances, the interactions between the considered factors, and a hierarchical structure of inputs and outputs.

Acknowledgments

The research of A. Labijak-Kowalska was supported by the Polish Ministry of Education and Science, grant no. 0311/SBAD/0709. M. Kadziński acknowledges financial support from the Polish National Science Center under the SONATA BIS project (grant no. DEC-2019/34/E/HS4/00045). L.C. Dias was supported by the Portuguese Foundation for Science and Technology (grant UIDB/05037/2020).

References

- Akkan, C., Karadayi, M.A., Ekinci, Y., Ülengin, F., Uray, N., Karaosmanoğlu, E., 2020. Efficiency analysis of emergency departments in metropolitan areas. *Socio-Economic Planning Sciences* 69, 100679.
- Amado, C., Santos, S., 2009. Challenges for performance assessment and improvement in primary health care: the case of the Portuguese health centres. *Health Policy* 91, 1, 43–56.
- Andes, S., Metzger, L.M., Kralewski, J., Gans, D., 2002. Measuring efficiency of physician practices using data envelopment analysis. *Managed Care* 11, 11, 48.
- 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.
- Basson, M.D., Butler, T., 2006. Evaluation of operating room suite efficiency in the veterans health administration system by using data envelopment analysis. *The American Journal of Surgery* 192, 5, 649–656.
- Charnes, A., Cooper, W., Golany, B., Seiford, L., Stutz, J., 1985. Foundations of data envelopment analysis for Pareto-Koopmans efficient empirical production functions. *Journal of Econometrics* 30, 1, 91–107.

- Charnes, A., Cooper, W., Wei, Q., Huang, Z., 1989. Cone-ratio data envelopment analysis and multi-objective programming. *International Journal of Systems Science* 20, 1099–1118.
- Charnes, A., Cooper, W.W., Rhodes, E., 1978. Measuring the efficiency of decision making units. *European Journal of Operational Research* 2, 6, 429–444.
- Chen, Y., Wang, J., Zhu, J., Sherman, H.D., Chou, S.Y., 2019. How the great recession affects performance: a case of Pennsylvania hospitals using DEA. *Annals of Operations Research* 278, 1–2, 77–99.
- Chilingerian, J.A., 1995. Evaluating physician efficiency in hospitals: a multivariate analysis of best practices. *European Journal of Operational Research* 80, 3, 548–574.
- Chilingerian, J.A., Sherman, H.D., 1990. Managing physician efficiency and effectiveness in providing hospital services. *Health Services Management Research* 3, 1, 3–15.
- Choi, S.O., 2005. Relative efficiency of fire and emergency services in Florida: an application and test of data envelopment analysis. *International Journal of Emergency Management* 2, 3, 218–230.
- Ciomek, K., Kadziński, M., 2021. Polyrun: a Java library for sampling from the bounded convex polytopes. *SoftwareX* 13, 100659.
- Contreras, I., 2020. A review of the literature on DEA models under common set of weights. *Journal of Modelling in Management* 15(4), 1277–1300.
- Cook, W., Roll, Y., Kazakov, A., 1990. A DEA model for measuring the relative efficiency of highway maintenance patrols. *INFOR* 28, 113–124.
- Färe, R., Grosskopf, S., 2000. Theory and application of directional distance functions. *Journal of Productivity Analysis* 13, 93–103.
- Fiallos, J., Patrick, J., Michalowski, W., Farion, K., 2017. Using data envelopment analysis for assessing the performance of pediatric emergency department physicians. *Health Care Management Science* 20, 1, 129–140.
- Flokou, A., Aletas, V., Niakas, D., 2017. A window-DEA based efficiency evaluation of the public hospital sector in Greece during the 5-year economic crisis. *PLoS ONE* 12, 5, e0177946.
- Gerami, J., Mavi, R.K., Saen, R.F., Mavi, N.K., 2020. A novel network DEA-R model for evaluating hospital services supply chain performance. *Annals of Operations Research* 1–26. <https://doi.org/10.1007/s10479-020-03755-w>.
- Goddard, M., Jacobs, R., 2009. Using composite indicators to measure performance in health care. In Smith, P.C., Mossialos, E., Leatherman, S., Papanicolas, I. (eds) *Performance Measurement for Health System Improvement: Experiences, Challenges and Prospects*. Cambridge University Press, Cambridge, pp. 339–368.
- Gouveia, M., Dias, L., Antunes, C., Mota, M., Duarte, E., Tenreiro, E., 2016. An application of value-based DEA to identify the best practices in primary health care. *OR Spectrum* 38, 3, 743–767.
- Gouveia, M.C., Dias, L.C., Antunes, C.H., 2008. Additive DEA based on MCDA with imprecise information. *Journal of the Operational Research Society* 59, 1, 54–63.
- Greco, S., Kadziński, M., Mousseau, V., Słowiński, R., 2012. Robust ordinal regression for multiple criteria group decision: UTA-GMS-GROUP and UTADIS-GMS-GROUP. *Decision Support Systems* 52, 3, 549–561.
- Greco, S., Mousseau, V., Słowiński, R., 2008. Ordinal regression revisited: Multiple criteria ranking using a set of additive value functions. *European Journal of Operational Research* 191, 2, 416–436.
- Hung, G., Chalut, D., 2008. A consensus-established set of important indicators of pediatric emergency department performance. 24, 9–15.
- Jacobs, R., Goddard, M., Smith, P.C., 2005. How robust are hospital ranks based on composite performance measures? *Medical Care* pp. 1177–1184.
- Jennings, N., Lee, G., Chao, K., Keating, S., 2009. A survey of patient satisfaction in a metropolitan emergency department: comparing nurse practitioners and emergency physicians. *International Journal of Nursing Practice* 15, 3, 213–218.
- Johannessen, K.A., Kittelsen, S.A., Hagen, T.P., 2017. Assessing physician productivity following Norwegian hospital reform: a panel and data envelopment analysis. *Social Science & Medicine* 175, 117–126.
- Kadziński, M., Greco, S., Słowiński, R., 2012a. Selection of a representative value function in robust multiple criteria ranking and choice. *European Journal of Operational Research* 217, 3, 541–553.
- Kadziński, M., Greco, S., Słowiński, R., 2012b. Extreme ranking analysis in robust ordinal regression. *Omega* 40, 3, 488–501.
- Kadziński, M., Labijak, A., Napieraj, M., 2017. Integrated framework for robustness analysis using ratio-based efficiency model with application to evaluation of Polish airports. *Omega* 67, 1–18.

- Kang, H., Nembhard, H., DeFlitch, C., Pasupathy, K., 2017. Assessment of emergency department efficiency using data envelopment analysis. *IISE Transactions on Healthcare Systems Engineering* 7, 4, 236–246.
- Keeney, R.L., Raiffa, H., 1993. *Decisions with Multiple Objectives: Preferences and Value Trade-Offs*. Cambridge University Press, Cambridge.
- Ketabi, S., Teymouri, E., Ketabi, M., 2018. Efficiency measurement of emergency departments in Isfahan, Iran. *International Journal of Process Management and Benchmarking* 8, 2, 142–155.
- Khushalani, J., Ozcan, Y.A., 2017. Are hospitals producing quality care efficiently? An analysis using dynamic network data envelopment analysis (DEA). *Socio-Economic Planning Sciences* 60, 15–23.
- Kohl, S., Schoenfelder, J., Fügener, A., Brunner, J.O., 2019. The use of data envelopment analysis (DEA) in healthcare with a focus on hospitals. *Health Care Management Science* 22, 2, 245–286.
- Kooreman, P., 1994. Nursing home care in The Netherlands: a nonparametric efficiency analysis. *Journal of Health Economics* 13, 3, 301–316.
- Küçük, A., Özsoy, V.S., Balkan, D., 2020. Assessment of technical efficiency of public hospitals in Turkey. *European Journal of Public Health* 30, 2, 230–235.
- Kuwahara, Y., Nagata, S., Taguchi, A., Naruse, T., Kawaguchi, H., Murashima, S., 2013. Measuring the efficiencies of visiting nurse service agencies using data envelopment analysis. *Health Care Management Science* 16, 3, 228–235.
- Labijak-Kowalska, A., Kadziński, M., 2021. Experimental comparison of results provided by ranking methods in data envelopment analysis. *Expert Systems with Applications* 170, 114739.
- Lahdelma, R., Salminen, P., 2006. Stochastic multicriteria acceptability analysis using the data envelopment model. *European Journal of Operational Research* 173, 1, 241–252.
- Lee, R.H., Bott, M.J., Gajewski, B., Taunton, R.L., 2009. Modeling efficiency at the process level: an examination of the care planning process in nursing homes. *Health Services Research* 44, 1, 15–32.
- Liu, J.S., Lu, L.Y., Lu, W.M., Lin, B.J., 2013. A survey of DEA applications. *Omega* 41, 5, 893–902.
- Meyer, P., Bigaret, S., 2012. Diviz: a software for modeling, processing and sharing algorithmic workflows in MCDA. *Intelligent Decision Technologies* 6, 4, 283–296.
- Ozcan, Y.A., Jiang, H., Pai, C.W., 2000. Do primary care physicians or specialists provide more efficient care? *Health Services Management Research* 13, 2, 90–96.
- Rouyendegh, B.D., Oztekin, A., Ekong, J., Dag, A., 2019. Measuring the efficiency of hospitals: a fully-ranking DEA–FAHP approach. *Annals of Operations Research* 278, 1, 361–378.
- Salo, A., Punkka, A., 2011. Ranking intervals and dominance relations for ratio-based efficiency analysis. *Management Science* 57, 1, 200–214.
- Schang, L., Hynninen, Y., Morton, A., Salo, A., 2016. Developing robust composite measures of healthcare quality-ranking intervals and dominance relations for Scottish Health Boards. *Social Science & Medicine* 162, 59–67.
- Shimshak, D.G., Lenard, M.L., Klimberg, R.K., 2009. Incorporating quality into data envelopment analysis of nursing home performance: a case study. *Omega* 37, 3, 672–685.
- Siddharthan, K., Ahern, M., Rosenman, R., 2000. Data envelopment analysis to determine efficiencies of health maintenance organizations. *Health Care Management Science* 3, 1, 23–29.
- Smith, C.A., Varkey, A.B., Evans, A.T., Reilly, B.M., 2004. Evaluating the performance of inpatient attending physicians. *Journal of General Internal Medicine* 19, 7, 766–771.
- Testi, A., Fareed, N., Ozcan, Y.A., Tanfani, E., 2013. Assessment of physician performance for diabetes: a bias-corrected data envelopment analysis model. *Quality in Primary Care* 21, 6, 345–357.
- Thau, M., Mikkelsen, M.F., Hjortskov, M., Pedersen, M.J., 2020. Question order bias revisited: a split-ballot experiment on satisfaction with public services among experienced and professional users. *Public Administration* 99, 189–204.
- Thompson, R., Langemeier, L., Lee, C., Lee, E., Thrall, R., 1990. The role of multiplier bounds in efficiency analysis with application to kansas farming. *Journal of Econometrics* 46, 93–108.
- Tosun, Ö., 2012. Using data envelopment analysis—neural network model to evaluate hospital efficiency. *International Journal of Productivity and Quality Management* 9, 2, 245–257.
- Veloso, A.S., Vaz, C.B., Alves, J., 2018. determinants of nursing homes performance: the case of portuguese santas casas da misericórdia. In Vaz, A.I.F., Almeida, J.P., Oliveira, J.F., Pinto, A.A. (eds) *Operational Research*. Springer, Cham, pp. 393–409.

- Wagner, J.M., Shimshak, D.G., 2000. Physician profiling using data envelopment analysis: a case study. *International Journal of Healthcare Technology and Management* 2, 1–4, 358–374.
- Winkler, R.L. and Hay, W.L., 1985. *Statistics: probability, inference, and decision*. Rinehart & Winston, New York.
- Zehra, Ö., Serpil, S., 2018. Evaluating healthcare system efficiency of OECD countries: a DEA-based study. In Kahraman, C., Ilker Topcu, Y. (eds) *Operations Research Applications in Health Care Management*. Springer, Cham, pp. 141–158.

Appendix

A.1. Computing the stochastic acceptability indices: an illustrative example

In this section, we discuss how to estimate the distribution of distances to the efficient unit and how to compute the ranks of physicians based on the expected efficiency, distance, or rank. We apply the hit-and-run algorithm to derive samples of weights for all inputs and outputs. Table A1 shows 10 examples of weight vectors used to compute the illustrative results in this section. Note that the outcomes reported in the main paper are derived from the analysis of 10,000 samples, which offers sufficient precision of the estimation.

Then, we compute a value-based efficiency score for each physician and each sample (see Table A2). When considering MD_i , its distance to the efficient unit is calculated as the difference between the maximal efficiency score of any physician obtained for a given sample and the efficiency score of MD_i . For example, for sample 1 and MD3, such a distance is equal to $d_3 = 0.275 - 0.066 = 0.209$. An efficiency rank of MD_i is computed based on the number of physicians with greater efficiencies than MD_i . For example, for sample 1, there are three physicians (MD6, MD11, and MD20) ranked better than MD5, and hence it is ranked fourth. The distances to the efficient unit and efficiency ranks for all physicians and samples are provided in Table A2.

Having computed the distances to the efficient unit for each decision-making unit (DMU) and each sample, we calculate $DAII$ as the ratio of the number of samples for which the distance lies within the analyzed interval to the number of all considered samples (see Table A3). For example, $DAII(MD1, (0.1, 0.2])$ is equal to 0.3 because for MD1, its distance to the efficient unit is in the $(0.1, 0.2]$ interval for 3 of 10 samples (samples 2, 5, and 9). The distributions of efficiency scores ($EAIIs$), ranks ($ERAIIs$), and preference relations ($PEOIs$) are computed analogously.

The results obtained for various samples can be averaged to estimate the expected measure values. The expected efficiencies EE , distances Ed , and ranks ER are presented in Table A2. To impose a

Table A1

Ten examples of input and output weight vectors obtained with the Monte Carlo simulation (for each vector, the weights sum up to 1)

	1	2	3	4	5	6	7	8	9	10
w_{i_1}	0.285	0.185	0.348	0.215	0.440	0.060	0.325	0.324	0.268	0.162
w_{i_2}	0.025	0.456	0.304	0.135	0.158	0.296	0.050	0.258	0.051	0.062
w_{i_3}	0.499	0.016	0.301	0.376	0.142	0.471	0.383	0.355	0.474	0.289
w_{o_1}	0.191	0.344	0.047	0.274	0.261	0.174	0.242	0.063	0.207	0.487

Table A2
Efficiency scores E , distances d , and ranks R for the considered physicians obtained and 10 examples of weight vectors

Sample	MD																				
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
1	E	0.216	0.083	0.066	0.116	0.252	0.275	0.134	0.146	0.172	0.064	0.262	0.109	0.218	0.126	0.203	0.177	0.126	0.175	0.199	0.261
	d	0.059	0.192	0.209	0.159	0.023	0.000	0.141	0.128	0.103	0.211	0.013	0.166	0.057	0.149	0.071	0.098	0.149	0.100	0.076	0.014
	R	6	18	19	16	4	1	13	12	11	20	2	17	5	15	7	9	14	10	8	3
2	E	0.435	0.281	0.243	0.417	0.527	0.611	0.419	0.438	0.464	0.401	0.595	0.411	0.582	0.378	0.484	0.480	0.406	0.536	0.471	0.547
	d	0.176	0.331	0.368	0.194	0.084	0.000	0.193	0.173	0.147	0.210	0.017	0.201	0.029	0.233	0.128	0.131	0.205	0.075	0.140	0.064
	R	12	19	20	14	6	1	13	11	10	17	2	15	3	18	7	8	16	5	9	4
3	E	0.122	0.188	0.188	0.310	0.478	0.462	0.315	0.359	0.359	0.281	0.411	0.283	0.394	0.305	0.411	0.322	0.309	0.346	0.418	0.476
	d	0.356	0.290	0.291	0.168	0.000	0.016	0.163	0.120	0.120	0.197	0.067	0.196	0.085	0.173	0.067	0.156	0.169	0.132	0.060	0.002
	R	20	18	19	13	1	3	12	9	8	17	5	16	7	15	6	11	14	10	4	2
4	E	0.314	0.137	0.112	0.192	0.309	0.358	0.207	0.213	0.243	0.150	0.354	0.192	0.316	0.183	0.265	0.267	0.199	0.276	0.255	0.321
	d	0.045	0.221	0.247	0.166	0.050	0.000	0.151	0.146	0.116	0.208	0.004	0.166	0.043	0.175	0.093	0.091	0.159	0.083	0.103	0.037
	R	5	19	20	15	6	1	13	12	11	18	2	16	4	17	9	8	14	7	10	3
5	E	0.312	0.155	0.109	0.223	0.384	0.460	0.226	0.278	0.302	0.175	0.412	0.191	0.395	0.243	0.339	0.242	0.202	0.323	0.335	0.419
	d	0.148	0.305	0.351	0.237	0.076	0.000	0.234	0.182	0.157	0.284	0.048	0.269	0.065	0.217	0.121	0.218	0.258	0.137	0.125	0.041
	R	9	19	20	15	5	1	14	11	10	18	3	17	4	12	6	13	16	8	7	2
6	E	0.236	0.187	0.193	0.290	0.390	0.384	0.308	0.292	0.309	0.268	0.394	0.304	0.352	0.246	0.334	0.381	0.315	0.338	0.328	0.374
	d	0.157	0.207	0.201	0.104	0.004	0.009	0.086	0.101	0.084	0.126	0.000	0.089	0.041	0.147	0.060	0.013	0.079	0.055	0.066	0.020
	R	18	20	19	15	2	3	12	14	11	16	1	13	6	17	8	4	10	7	9	5
7	E	0.272	0.102	0.072	0.139	0.278	0.327	0.153	0.175	0.203	0.087	0.307	0.126	0.270	0.152	0.233	0.192	0.139	0.218	0.227	0.298
	d	0.055	0.225	0.255	0.188	0.049	0.000	0.174	0.152	0.124	0.240	0.020	0.201	0.057	0.175	0.094	0.135	0.188	0.109	0.100	0.029
	R	5	18	20	16	4	1	13	12	10	19	2	17	6	14	7	11	15	9	8	3
8	E	0.129	0.170	0.169	0.277	0.439	0.424	0.285	0.321	0.325	0.245	0.381	0.255	0.358	0.273	0.374	0.300	0.280	0.313	0.379	0.436
	d	0.309	0.269	0.270	0.161	0.000	0.015	0.153	0.118	0.114	0.194	0.058	0.184	0.081	0.166	0.065	0.139	0.158	0.125	0.060	0.003
	R	20	18	19	14	1	3	12	9	8	17	4	16	7	15	6	11	13	10	5	2
9	E	0.235	0.096	0.077	0.134	0.265	0.293	0.151	0.162	0.188	0.084	0.282	0.129	0.239	0.139	0.218	0.198	0.144	0.197	0.212	0.275
	d	0.058	0.198	0.216	0.159	0.028	0.000	0.142	0.131	0.105	0.209	0.011	0.164	0.054	0.154	0.075	0.095	0.150	0.096	0.081	0.018
	R	6	18	20	16	4	1	13	12	11	19	2	17	5	15	7	9	14	10	8	3
10	E	0.509	0.128	0.072	0.145	0.229	0.353	0.163	0.152	0.206	0.096	0.374	0.158	0.329	0.136	0.206	0.256	0.147	0.284	0.183	0.264
	d	0.000	0.381	0.438	0.364	0.281	0.156	0.347	0.357	0.303	0.414	0.135	0.351	0.180	0.373	0.303	0.253	0.363	0.225	0.327	0.246
	R	1	18	20	16	8	3	12	14	9	19	2	13	4	17	10	7	15	5	11	6
EE	0.278	0.153	0.130	0.224	0.355	0.395	0.236	0.254	0.277	0.185	0.377	0.216	0.345	0.218	0.307	0.282	0.227	0.301	0.301	0.367	
Ed	0.136	0.262	0.284	0.190	0.059	0.020	0.178	0.161	0.137	0.229	0.037	0.199	0.069	0.196	0.108	0.133	0.188	0.114	0.114	0.047	
ER	10.2	18.5	19.6	15.0	4.1	1.8	12.7	11.6	9.9	18.0	2.5	15.7	5.1	15.5	7.3	9.1	14.1	8.1	7.9	3.3	

These results are used to estimate the expected efficiencies EE , distances Ed , and ranks ER .

Table A3

Distribution of the distances to the efficient unit (*DAIIs*) based on 10 examples of weight vectors

MD	[0.0, 0.1]	(0.1, 0.2]	(0.2, 0.3]	(0.3, 0.4]	(0.4, 0.5]	(0.5, 0.6]	(0.6, 0.7]	(0.7, 0.8]	(0.8, 0.9]	(0.9, 1.0]
1	0.5	0.3	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.2	0.5	0.3	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.7	0.2	0.1	0.0	0.0	0.0	0.0	0.0
4	0.0	0.8	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0
5	0.9	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
6	0.9	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
7	0.1	0.7	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0
8	0.0	0.9	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0
9	0.1	0.8	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0
10	0.0	0.3	0.6	0.0	0.1	0.0	0.0	0.0	0.0	0.0
11	0.9	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
12	0.1	0.5	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0
13	0.9	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
14	0.0	0.7	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0
15	0.7	0.2	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0
16	0.4	0.4	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
17	0.1	0.6	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0
18	0.4	0.5	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
19	0.5	0.4	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0
20	0.9	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0

complete order on the set of physicians, they need to be sorted accordingly (e.g., in the ascending order when accounting for *Ed*). In the considered example, the best (minimal) expected distance is associated with MD6 (0.020) and the worst (maximal) distance is attained by MD3 (0.284). These physicians are ranked at top and bottom, respectively. The rankings based on the expected ranks (*ERs*) or efficiencies can be constructed analogously. Note, however, that while lower distances and ranks are preferred, greater values are more favorable when considering the efficiency scores.

A.2. Descriptive statistics of input and output data for the three considered complaint groups

In Table A4, we report the descriptive statistics of input and output data for the three complaint groups considered in the main paper: *G1*—abdominal pain and constipation; *G2*—fever; and *G3*—lower or upper extremity injury, head injury, and laceration/puncture.

A.3. Distributions of the distances to the efficient unit and the efficiency scores for complaint group *G1*

In Tables A5 and A6, we report the distributions of the distances to the efficient unit and the efficiency scores for complaint group *G1* estimated based on 10,000 weight vectors. They are captured

Table A4

Descriptive statistics of input and output data for the three considered complaint groups ($G1$, $G2$, and $G3$)

Group	Statistic	i_1 – AVG_MDTIME_PAT	i_2 – AVG_LAB_PAT	i_3 – AVG_RAD_PAT	o_1 – RATE_NR72
$G1$	Min	1.435	0.857	0.462	0.905
	Max	2.311	2.760	0.920	1.000
	Mean	1.811	1.739	0.656	0.958
	St. dev. SD	0.254	0.431	0.112	0.022
$G2$	Min	1.017	0.357	0.207	0.907
	Max	1.752	1.101	0.419	1.000
	Mean	1.367	0.668	0.322	0.963
	SD	0.227	0.206	0.061	0.020
$G2$	Min	0.836	0.000	0.478	0.957
	Max	1.293	0.176	0.847	1.000
	Mean	1.058	0.071	0.684	0.985
	SD	0.132	0.055	0.090	0.010

Table A5

Distribution of the distances to the efficient unit ($DAIIs$)

MD	[0.0, 0.1]	(0.1, 0.2]	(0.2, 0.3]	(0.3, 0.4]	(0.4, 0.5]	(0.5, 0.6]	(0.6, 0.7]	(0.7, 0.8]	(0.8, 0.9]	(0.9, 1.0]
1	0.405	0.237	0.187	0.120	0.051	0.000	0.000	0.000	0.000	0.000
2	0.000	0.017	0.617	0.365	0.001	0.000	0.000	0.000	0.000	0.000
3	0.000	0.005	0.423	0.518	0.054	0.000	0.000	0.000	0.000	0.000
4	0.004	0.598	0.344	0.054	0.000	0.000	0.000	0.000	0.000	0.000
5	0.750	0.196	0.050	0.004	0.000	0.000	0.000	0.000	0.000	0.000
6	0.955	0.043	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000
7	0.027	0.618	0.309	0.046	0.000	0.000	0.000	0.000	0.000	0.000
8	0.008	0.781	0.170	0.041	0.000	0.000	0.000	0.000	0.000	0.000
9	0.072	0.834	0.085	0.009	0.000	0.000	0.000	0.000	0.000	0.000
10	0.000	0.365	0.493	0.122	0.020	0.000	0.000	0.000	0.000	0.000
11	0.965	0.035	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
12	0.022	0.470	0.434	0.074	0.000	0.000	0.000	0.000	0.000	0.000
13	0.891	0.103	0.006	0.000	0.000	0.000	0.000	0.000	0.000	0.000
14	0.000	0.515	0.420	0.063	0.002	0.000	0.000	0.000	0.000	0.000
15	0.495	0.420	0.076	0.009	0.000	0.000	0.000	0.000	0.000	0.000
16	0.295	0.512	0.191	0.002	0.000	0.000	0.000	0.000	0.000	0.000
17	0.044	0.511	0.373	0.071	0.001	0.000	0.000	0.000	0.000	0.000
18	0.418	0.553	0.029	0.000	0.000	0.000	0.000	0.000	0.000	0.000
19	0.457	0.432	0.090	0.021	0.000	0.000	0.000	0.000	0.000	0.000
20	0.853	0.119	0.028	0.000	0.000	0.000	0.000	0.000	0.000	0.000

by distance acceptability interval indices $DAIIs$, and efficiency acceptability interval indices, $EAIIs$, respectively. These results are referred to in Section 3.1 of the main paper.

The analysis of such distributions allows identifying the DMUs for which the results vary much in the set of feasible weights. High dispersion of scores and distances should prompt investigation as to whether the guidelines for standard practice can be used to reduce variance in management.

Table A6
Distribution of the efficiency scores (*EAI*s)

MD	[0.0, 0.1]	(0.1, 0.2]	(0.2, 0.3]	(0.3, 0.4]	(0.4, 0.5]	(0.5, 0.6]	(0.6, 0.7]	(0.7, 0.8]	(0.8, 0.9]	(0.9, 1.0]
1	0.030	0.183	0.230	0.265	0.206	0.086	0.000	0.000	0.000	0.000
2	0.040	0.533	0.427	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3	0.199	0.485	0.316	0.000	0.000	0.000	0.000	0.000	0.000	0.000
4	0.000	0.233	0.332	0.325	0.110	0.000	0.000	0.000	0.000	0.000
5	0.000	0.012	0.152	0.318	0.320	0.187	0.011	0.000	0.000	0.000
6	0.000	0.000	0.030	0.250	0.399	0.291	0.030	0.000	0.000	0.000
7	0.000	0.197	0.348	0.335	0.120	0.000	0.000	0.000	0.000	0.000
8	0.001	0.150	0.335	0.321	0.188	0.005	0.000	0.000	0.000	0.000
9	0.000	0.063	0.324	0.371	0.239	0.003	0.000	0.000	0.000	0.000
10	0.098	0.274	0.282	0.277	0.069	0.000	0.000	0.000	0.000	0.000
11	0.000	0.000	0.032	0.313	0.402	0.244	0.009	0.000	0.000	0.000
12	0.019	0.243	0.341	0.312	0.085	0.000	0.000	0.000	0.000	0.000
13	0.000	0.001	0.118	0.348	0.355	0.175	0.003	0.000	0.000	0.000
14	0.009	0.216	0.423	0.332	0.020	0.000	0.000	0.000	0.000	0.000
15	0.000	0.029	0.253	0.369	0.309	0.040	0.000	0.000	0.000	0.000
16	0.000	0.075	0.304	0.331	0.267	0.023	0.000	0.000	0.000	0.000
17	0.004	0.236	0.332	0.325	0.103	0.000	0.000	0.000	0.000	0.000
18	0.000	0.018	0.231	0.380	0.307	0.064	0.000	0.000	0.000	0.000
19	0.000	0.056	0.262	0.355	0.285	0.042	0.000	0.000	0.000	0.000
20	0.000	0.006	0.099	0.322	0.364	0.196	0.013	0.000	0.000	0.000

In our study, the example units for which such verification should be carried out are MD1, MD8, MD12, MD17, and MD19.

A.4. Analysis of efficiency ranks for complaint group G1

In this section, we discuss the robustness of efficiency ranks for complaint group *G1*. The distances to the efficient DMU and efficiency scores are derived from the cardinal-oriented comparison of physicians. In turn, efficiency ranks build on the ordinal comparisons between the physicians. In Table A7, we report the extreme (R_* and R^*) and expected (ER) ranks. The physicians identified as efficient have the best ranks equal to 1. Based on R_* , MD13 is the best among the inefficient units. (S)he is ranked second in the best case $R_* = 2$), which means that in the most favorable scenario, it is less efficient only than a single efficient MD, while attaining better scores than the remaining 18 physicians. MD2 and MD3 have the least positive results in terms of R_* . For these physicians, there are at least 13 and 17 other physicians in a group who are more efficient for any feasible weight vector.

The analysis of the worst efficiency ranks (R^*) indicates that four efficient physicians (MD5, MD6, MD11, and MD20) never fall out of the top eight. Thus, the stability of derived ranks is the highest for MD6 because even in the least favorable scenario, only four other physicians attain better efficiencies. The performance of the other two efficient physicians is less stable. In particular,

Table A7
The extreme and expected ranks, and efficiency rank acceptability indices (ERAI)s for the considered physicians

MD	R*	ER	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
1	1	20	11.041	0.166	0.040	0.055	0.039	0.028	0.039	0.023	0.045	0.031	0.017	0.037	0.028	0.005	0.014	0.011	0.024	0.028	0.138	0.023	0.209
2	14	20	18.669	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.003	0.013	0.017	0.261	0.691	0.015	
3	18	20	19.640	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.121	0.118	0.761	
4	11	18	14.409	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.026	0.069	0.081	0.270	0.405	0.146	0.003	0.000	0.000	0.000	
5	1	8	4.541	0.145	0.065	0.179	0.077	0.147	0.136	0.138	0.113	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
6	1	5	1.860	0.466	0.228	0.290	0.012	0.004	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
7	9	15	12.798	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.001	0.024	0.335	0.459	0.166	0.013	0.000	0.000	0.000	0.000	
8	8	17	11.693	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.056	0.094	0.138	0.209	0.253	0.045	0.066	0.055	0.078	0.006	0.000	0.000	
9	8	16	9.722	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.105	0.275	0.465	0.120	0.021	0.012	0.001	0.000	0.000	0.000	0.000	0.000	
10	13	20	17.347	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.004	0.037	0.231	0.262	0.283	0.168	0.015	
11	1	8	2.914	0.148	0.344	0.153	0.237	0.046	0.063	0.009	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
12	9	18	15.150	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.003	0.071	0.041	0.073	0.126	0.170	0.195	0.314	0.007	0.000	0.000	
13	2	14	4.763	0.000	0.000	0.181	0.328	0.256	0.058	0.142	0.031	0.002	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
14	10	19	15.339	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.088	0.057	0.133	0.080	0.089	0.064	0.307	0.180	0.000	
15	5	11	7.412	0.000	0.000	0.000	0.000	0.069	0.212	0.244	0.227	0.210	0.037	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
16	1	16	9.110	0.000	0.015	0.027	0.034	0.065	0.060	0.112	0.119	0.079	0.077	0.194	0.099	0.054	0.044	0.015	0.006	0.000	0.000	0.000	
17	6	18	14.347	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.004	0.008	0.031	0.035	0.077	0.122	0.206	0.201	0.243	0.063	0.010	0.000	
18	4	14	7.385	0.000	0.000	0.000	0.034	0.117	0.238	0.138	0.198	0.111	0.126	0.031	0.005	0.002	0.000	0.000	0.000	0.000	0.000	0.000	
19	4	15	8.178	0.000	0.000	0.000	0.099	0.086	0.107	0.104	0.099	0.188	0.101	0.164	0.015	0.014	0.023	0.000	0.000	0.000	0.000	0.000	
20	1	8	3.682	0.075	0.308	0.115	0.140	0.182	0.087	0.090	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	

MD16 is ranked 16th in the worst case, whereas MD1 is ranked at the very bottom. There are only three other physicians (MD2, MD3, and MD10) ranked 20th for at least one feasible weight vector.

The analysis of extreme efficiency ranks can be enriched with consideration of the *ERAI*s (see Table A7), indicating for each physician the distribution of ranks over the feasible weight vectors. For some physicians, the derived ranks are relatively stable. For example, MD3 is ranked at the very bottom for 76.1% weights and MD2 is ranked 18th or 19th for 95.2% samples. MD6 is ranked at the top for 46.6% weight, making him/her the most efficient physician in the group. In general, such a high value for the first rank acceptability index may indicate the outlier DMU. It may motivate the management to investigate the results without considering such an overall good performer who influences the distances of many other DMUs.

As far as MD13 is concerned, its possible rank interval is relatively wide [2, 14]. However, for 96.5% feasible weights, it is ranked in the top seven. For some other physicians, the ranks are more distributed. In particular, the *ERAI*s for MD1 are positive for all ranks with $ERAI(MD1, 1)$ (16.6%) being close to $ERAI(MD1, 20)$ (20.9%). This means that, depending on the chosen input/output weights, it is almost equally likely for MD1 to be ranked at the top or at the bottom. A similar distribution of ranks can be observed for MD16. For this physician, *ERAI*s are nonzero for ranks between 2 and 16, with the greatest one being lower than 0.2.

The *ER*s (see Table A7) can also be used to order all physicians. The top-ranked physicians are MD6 ($ER = 1.860$) and MD11 ($ER = 2.914$), whereas the bottom-ranked physicians are MD2 ($ER = 18.669$) and MD3 ($ER = 19.640$). The ranking determined by *ER*s is very similar to the orders imposed by *Eds* and *EEs*. The swaps occur only for two pairs, (MD5, MD13) and (MD17, MD4), which confirms the stability of conclusions derived from the multiperspective robustness analysis.

In general, the expected results exhibit which units perform good or bad for different priorities assigned to inputs and outputs. In some situations, the expected efficiencies or ranks of inefficient units can be, on average, better than for some efficient units (see, e.g., the average ranks of inefficient MD13 and MD15 compared to the expected positions for the efficient MD1 and MD16). Such results may indicate the need to implement the corrective actions for the average bad performers who prove to be efficient only under specific scenarios.

A.5. Analysis of pairwise preference relations for complaint group G1

Another aspect considered in the robustness analysis concerns pairwise comparisons between physicians. The Hasse diagram of the necessary preference relation is presented in Fig. A1. No physician is necessarily preferred over the six efficient physicians. However, there is also one inefficient physician (MD13) who is not necessarily worse than any other physician (depending on the weights, the physicians performing better than MD13 are not the same). Overall, MD5, MD6, and MD20 are necessarily preferred to the largest number of other physicians (12), which confirms their superior performance. On the other hand, MD1, MD2, MD3, and MD10 are not necessarily preferred to any other physician. MD1 can be seen as a potential outlier because it is neither necessarily better nor worse than any other physician.

The graph of the necessary preference relation can be used for constructing the corrective actions and improvement paths for inefficient physicians. From a short-term perspective, one can focus on

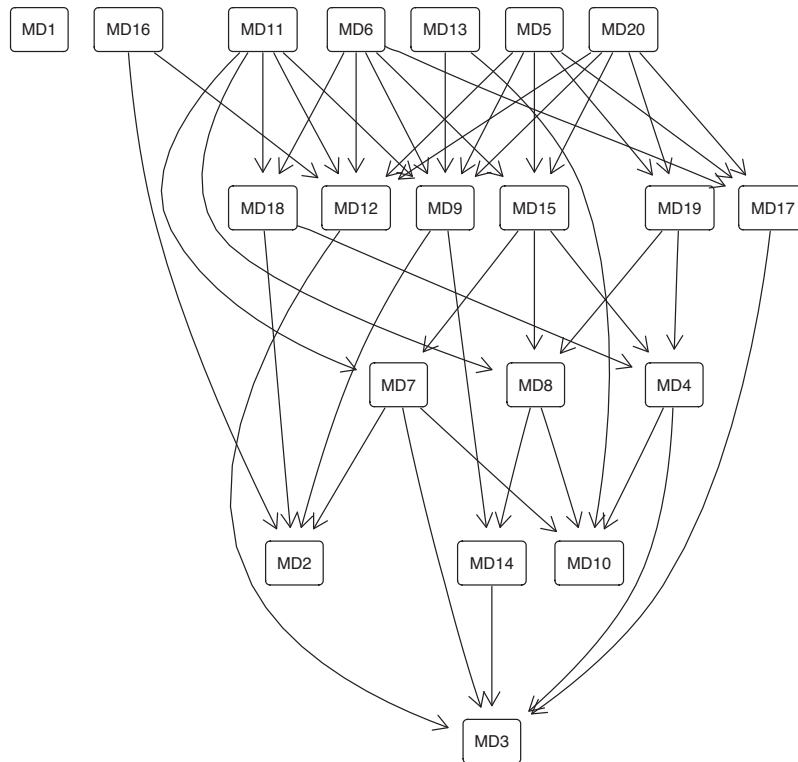


Fig. A1. The Hasse diagram of the necessary efficiency preference relation \succsim_E^N .

the units that are necessarily preferred over some inefficient DMUs. For example, for MD8, these can be MD11, MD15, or MD19. The differences in inputs and outputs for such units indicate the improvement potential. From a long-term perspective, one can apply the stepwise benchmarking based on the paths observed in the Hasse diagram of \succsim_E^N . For example, MD3—ranked at the bottom—can improve by following some improvement paths, for example, (MD14, MD8, MD19, MD5) or (MD7, MD15, MD20).

For pairs of physicians who are incomparable in terms of \succsim_E^N , the efficiency comparison results are not univocal, given all feasible weights. Such pairs are not connected by an arc in Fig. A1. The shares of feasible weights confirming one physician's better performance over another are captured by *PEOIs* (see Table A8). For some other pairs, one physician performs clearly better, for example, $PEOI(MD16, MD17) = 0.980$ indicates that for 98% of feasible weights, MD16 is at least as efficient as MD17. Thus, even if the preference relation is not fully robust for this pair, it is close to being so. Similar conclusions can be drawn for (MD18, MD12), (MD13, MD7), and (MD8, MD2). For some pairs of physicians these shares are more balanced, for example, for (MD13, MD5)— $PEOI(MD13, MD5) = 0.513$ and $PEOI(MD5, MD13) = 0.487$. Similar observations apply to (MD17, MD4) or (MD18, MD15).

The remaining DMUs do not influence such pairwise comparisons. The analyst may be interested in such a one-on-one perspective if (s)he knows some units better than others. Then, they can be

Table A8
Pairwise efficiency outranking indices (PEOIs) for all pairs of physicians

MD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	1.000	0.779	0.794	0.563	0.357	0.209	0.538	0.503	0.463	0.624	0.215	0.579	0.272	0.580	0.424	0.440	0.560	0.377	0.437	0.324
2	0.221	1.000	0.883	0.000	0.000	0.000	0.000	0.002	0.000	0.182	0.000	0.013	0.000	0.013	0.000	0.000	0.013	0.000	0.000	0.000
3	0.206	0.117	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.012	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
4	0.437	1.000	1.000	1.000	0.000	0.000	0.104	0.104	0.000	1.000	0.000	0.664	0.000	0.679	0.000	0.072	0.550	0.000	0.000	0.000
5	0.643	1.000	1.000	1.000	1.000	0.191	1.000	1.000	1.000	1.000	0.377	1.000	0.487	1.000	1.000	0.817	1.000	0.701	1.000	0.212
6	0.791	1.000	1.000	1.000	1.000	0.809	1.000	1.000	1.000	1.000	0.727	1.000	1.000	1.000	1.000	0.982	1.000	1.000	1.000	0.806
7	0.462	1.000	1.000	0.896	0.000	0.000	1.000	0.262	0.019	1.000	0.000	0.879	0.001	0.753	0.000	0.088	0.873	0.004	0.038	0.000
8	0.497	0.998	1.000	0.896	0.000	0.000	0.738	1.000	0.089	1.000	0.000	0.792	0.000	1.000	0.000	0.349	0.791	0.150	0.000	0.000
9	0.537	1.000	1.000	1.000	0.000	0.000	0.981	0.911	1.000	1.000	0.000	0.979	0.000	1.000	0.034	0.485	0.955	0.179	0.265	0.000
10	0.376	0.818	0.988	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.140	0.000	0.313	0.000	0.006	0.007	0.000	0.000	0.000
11	0.785	1.000	1.000	1.000	0.623	0.273	1.000	1.000	1.000	1.000	1.000	0.986	1.000	0.986	0.916	0.988	1.000	1.000	0.888	0.571
12	0.421	0.987	1.000	0.336	0.000	0.000	0.121	0.208	0.021	0.860	0.000	1.000	0.000	0.527	0.000	0.000	0.326	0.001	0.055	0.000
13	0.728	1.000	1.000	1.000	0.513	0.000	0.999	1.000	1.000	1.000	0.014	1.000	1.000	1.000	0.817	0.918	0.996	1.000	0.810	0.441
14	0.420	0.987	1.000	0.321	0.000	0.000	0.247	0.000	0.000	0.687	0.000	0.473	0.000	1.000	0.000	0.134	0.392	0.007	0.000	0.000
15	0.576	1.000	1.000	1.000	0.000	0.000	1.000	1.000	0.966	1.000	0.084	1.000	0.183	1.000	1.000	0.636	1.000	0.458	0.706	0.000
16	0.560	1.000	1.000	0.928	0.183	0.018	0.912	0.651	0.515	0.994	0.012	1.000	0.082	0.866	0.364	1.000	0.980	0.221	0.409	0.143
17	0.440	0.987	1.000	0.450	0.000	0.000	0.127	0.209	0.045	0.993	0.000	0.674	0.004	0.608	0.000	0.020	1.000	0.016	0.030	0.000
18	0.623	1.000	1.000	1.000	0.299	0.000	0.996	0.850	0.821	1.000	0.000	0.999	0.000	0.993	0.542	0.779	0.984	1.000	0.571	0.176
19	0.563	1.000	1.000	1.000	0.000	0.000	0.962	1.000	0.735	1.000	0.112	0.945	0.190	1.000	0.294	0.591	0.970	0.429	1.000	0.000
20	0.676	1.000	1.000	1.000	0.788	0.194	1.000	1.000	1.000	1.000	0.429	1.000	0.559	1.000	1.000	0.857	1.000	0.824	1.000	1.000

Table A9

Input and output values for the complaint groups $G2$ (fever) and $G3$ (lower or upper extremity injury, head injury, and laceration/puncture) by physician

Group MD	$G2$				$G3$			
	i_1	i_2	i_3	o_1	i_1	i_2	i_3	o_1
MD1	1.639	0.604	0.333	1.000	1.293	0.000	0.699	0.957
MD2	1.682	1.031	0.374	0.969	1.287	0.166	0.847	0.983
MD3	1.386	0.551	0.318	0.907	1.123	0.030	0.723	0.970
MD4	1.482	0.600	0.419	0.943	1.122	0.115	0.803	1.000
MD5	1.362	0.561	0.305	0.952	1.050	0.021	0.609	0.979
MD6	1.017	0.496	0.207	0.953	0.914	0.021	0.689	0.992
MD7	1.457	0.934	0.316	0.969	1.056	0.108	0.652	0.990
MD8	1.084	0.632	0.212	0.964	0.95	0.000	0.728	0.981
MD9	1.223	0.751	0.279	0.959	1.027	0.090	0.754	0.983
MD10	1.140	0.357	0.260	0.959	1.173	0.024	0.778	0.986
MD11	1.538	0.384	0.299	0.943	1.046	0.020	0.654	0.986
MD12	1.061	0.407	0.407	0.966	0.943	0.074	0.595	0.992
MD13	1.255	0.730	0.340	0.977	0.995	0.052	0.617	0.991
MD14	1.473	0.659	0.388	0.976	1.139	0.176	0.617	0.991
MD15	1.265	0.581	0.372	0.977	0.852	0.090	0.639	0.976
MD16	1.752	0.912	0.412	0.985	0.988	0.000	0.478	1.000
MD17	1.571	1.101	0.314	0.977	1.092	0.127	0.756	0.991
MD18	1.597	0.772	0.308	0.965	1.264	0.110	0.793	0.984
MD19	1.306	0.743	0.273	0.97	1.010	0.109	0.592	0.990
MD20	1.044	0.549	0.302	0.941	0.836	0.085	0.667	0.977

employed as fixed benchmarks for the inefficient DMUs. For example, if an expert knows MD16 quite well, (s)he may use it to formulate guidelines for MD2 and MD12, which are worse than MD16 for all possible weights assigned to inputs and outputs.

A.6. Input and output values for the complaint groups $G2$ and $G3$

In Table A9, we present the input and output values for the complaint groups $G2$ (fever) and $G3$ (lower or upper extremity injury, head injury, and laceration/puncture). Together with group $G1$, they form the basis for conducting a multiscenario robustness analysis, whose results are discussed in Section 3.3 of the main paper and Section A7.

A.7. The analysis of pairwise preference relations for a multiscenario setting

This section presents the pairwise comparisons of physicians for three complaint groups. Table A10 reports the truth of the necessary-necessary $\succsim_{E,S}^{N,N}$ and necessary-possible $\succsim_{E,S}^{N,P}$ preference relations for all pairs of physicians. Since $\succsim_{E,S}^{N,N}$ is transitive, it can be presented graphically by its Hasse diagram (see Fig. A2). For 10 pairs of physicians, the necessary preference relation holds for all complaint groups. In particular, five physicians (MD5, MD6, MD8, MD19, and MD20) are always

Table A10

The truth of the necessary-necessary $\succsim_{E,S}^{N,N}$ (NN) and necessary-possible $\succsim_{E,S}^{N,P}$ (NP) efficiency preference relations for all pairs of physicians based on the analysis of three complaint groups

MD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
1	NN	NP	NP																		
2		NN																			
3			NN																		
4		NP	NP	NN						NP										NP	
5	NP	NP	NN	NP	NN		NP	NP	NP	NP		NP		NP	NP		NP	NP	NP	NP	
6	NP	NP	NN	NP		NN	NP	NP	NN	NP		NP		NP	NP		NP	NP		NP	
7	NP	NN	NP				NN			NP				NP						NP	
8		NP	NN					NN		NP				NP							
9		NP	NP						NN					NP						NP	
10		NP	NP			NP		NP		NN										NP	
11	NP	NP	NP	NP			NP	NP	NP	NP	NN	NP		NP						NP	
12	NP	NP	NP		NP		NP		NP	NP	NP	NN	NP	NP		NP	NP	NP	NP	NP	
13	NP	NN	NP				NP	NP	NP	NP			NN	NN		NP	NP	NP		NP	
14	NP	NP	NP							NP				NN						NP	
15	NP	NP	NP	NP		NP	NP	NP		NP	NP			NP	NN					NP	NP
16	NP	NP	NP	NP	NP		NP		NP	NP	NP	NP	NP	NP		NN	NP	NP	NP	NP	
17		NP	NP											NP			NN	NP			
18		NN	NP	NP						NP										NN	
19	NP	NP	NN	NP				NP	NP	NP	NP			NP						NP	NN
20	NP	NP	NN	NP			NP	NP	NP	NP		NP		NP	NP				NP	NP	NN

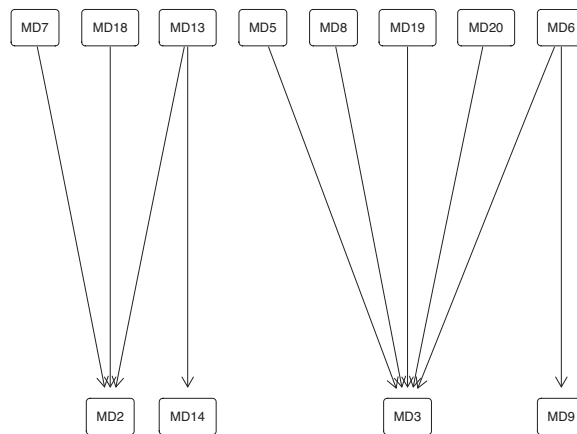


Fig. A2. The Hasse diagram of the necessary-necessary efficiency preference relation $\succsim_{E,S}^{N,N}$ based on the analysis of three complaint groups (for clarity of presentation, physicians not related by $\succsim_{E,S}^{N,N}$ with any other physician have been omitted).

Table A11

Ranks attained by physicians in the orders imposed by different measures derived from robustness analysis for complaint group $G1$ and differences between extreme distances, efficiencies, and ranks

MD	Ranks according to different measures											Widths of intervals		
	d_*	d^*	Ed	E^*	E_*	EE	R_*	R^*	ER	$ \succsim^N $	$ \prec^N $	$d^* - d_*$	$E^* - E_*$	$R^* - R_*$
1	1	20	11	8	19	11	1	17	11	17	1	0.558	0.560	19
2	19	11	19	19	17	19	19	17	19	17	18	0.227	0.250	6
3	20	19	20	20	20	20	20	17	20	17	20	0.306	0.265	2
4	15	13	14	15	12	14	17	13	15	13	16	0.323	0.365	7
5	1	7	5	2	5	5	1	2	4	1	1	0.366	0.522	7
6	1	2	1	3	1	1	1	1	1	1	1	0.260	0.408	4
7	13	10	13	17	9	13	14	8	13	9	12	0.319	0.330	6
8	16	15	12	11	14	12	12	12	12	9	15	0.340	0.461	9
9	14	8	10	12	6	10	12	10	10	9	12	0.293	0.387	8
10	17	18	18	18	18	18	18	17	18	17	19	0.368	0.397	7
11	1	1	2	4	2	2	1	2	2	4	1	0.204	0.429	7
12	12	14	16	13	15	16	14	13	16	14	12	0.354	0.393	9
13	7	3	4	5	3	4	7	6	5	6	1	0.254	0.438	12
14	18	17	17	16	16	17	16	16	17	14	17	0.304	0.391	9
15	9	9	7	9	8	7	10	5	7	5	10	0.332	0.467	6
16	1	6	9	10	10	9	1	10	9	9	1	0.361	0.455	15
17	11	16	15	14	13	15	11	13	14	14	10	0.372	0.368	12
18	8	4	6	7	7	6	8	6	6	8	8	0.252	0.469	10
19	10	12	8	6	11	8	8	8	8	6	8	0.358	0.509	11
20	1	5	3	1	4	3	1	2	3	1	1	0.346	0.511	7

at least as efficient as MD3, and three physicians (MD7, MD18, and MD13) are more efficient than MD2. MD13 and MD6 can serve as the benchmark to follow for two other pairs (MD2 and MD14 or MD3 and MD9, respectively).

The necessary-possible preference relation $\succsim_{E,S}^{N,P}$ is more dense (see Table A10; note that the truth of $\succsim_{E,S}^{N,N}$ implies $\succsim_{E,S}^{N,P}$). There are 153 ordered pairs of physicians for whom the necessary relation holds for at least one complaint group. Interestingly, for some pairs (e.g., MD10, MD18), this relation is instantiated in both directions. Such observations, along with a high density of $\succsim_{E,S}^{N,P}$ and a scarcity of $\succsim_{E,S}^{N,N}$, suggest that the performance of physicians is strongly related to the complaint group and therefore some of them are better in treating specific groups of patients.

A.8. Summary of results derived from the robustness analysis

In this section, we summarize the results derived for each physician with the proposed robustness analysis framework for a single scenario referring to complaint group $G1$ and multiple scenarios concerning groups $G1$ – $G3$.

In Table A11, we present the ranks of all physicians in the orders imposed by different measures following the application of robust efficiency analysis framework to group $G1$. These measures include extreme and expected distances, efficiency score, and ranks as well as the numbers of other

physicians which are less ($|\succsim^N|$) or more ($|\precsim^N|$) preferred than a given physician according to the necessary relation. The rankings are enriched with the differences between extreme distances, efficiencies, and ranks that indicate the stability of results for each physician.

These results confirm that the efficiency results are stable for some physicians irrespective of the accounted perspective and considered weight vectors. For example, MD6 is ranked at the top for 9 of 11 considered measures while attaining the second and third positions in the rankings determined by d^* and E^* , respectively. Such favorable results are justified by the relatively good performance of MD6 on all inputs and outputs. Furthermore, MD11 and MD20 also attain the ranks among the top five MDs according to all measures. On the other extreme, MD2, MD3, MD10, and MD14 are ranked relatively low. For example, MD3 is never ranked better than 17th. Its scores, efficiencies, and ranks are stable irrespective of the considered weights with the interval widths equal to 0.306, 0.264, and 2, respectively. This is understandable given its unfavorable performances on all accounted factors.

Even though the ranks attained by the vast majority of physicians are relatively stable irrespective of the accounted measure, one can indicate a few examples for which these indications are inconsistent. This is because of their unbalanced input/output profiles, making their performance strongly dependent on the considered weights and their ranks more prone to fluctuations with the change in the accounted measure. For example, the widest distance, efficiency, and rank intervals can be observed for MD1. Its ranks range from the most favorable (see, e.g., d_* and R_*) through medium (see, e.g., Ed , EE , and ER) to the least favorable (see, e.g., d^* and E_*). The great variability of results can also be noted for MD16. Its rank ranges from first (see, e.g., d_* and R_*) to tenth (see, e.g., E^* and E_*) depending on the selected measure, whereas a difference between extreme efficiency ranks ($R^* - R_*$) is 15.

Analogous results derived from the analysis of three complaint groups are presented in Table A12. The considered measures are extreme possible-possible distances to the efficient physician, efficiency scores, and ranks as well as the numbers of physicians which proved to be worse ($|\succsim^{N,P}|$) or better ($|\precsim^{N,P}|$) than a given physician according to the necessary-possible relation.

The ranks attained by different physicians according to the measures quantifying the results for multiple scenarios are, in general, less stable than for a single complaint group only. This confirms that the considered physicians attain more favorable results for complaint groups for which they have specialized skills while performing worse for other groups. Nevertheless, the conclusions on the best and worse performing physicians are similar. For example, MD15 attains ranks between first (see d_* and R_*) and eighth (see E_*) in the orders imposed by different measures. Furthermore, when considering the numbers of other physicians who proved to be necessarily-possibly worse or better than MD15, it is ranked sixth. Also, MD3 attains relatively stable ranks. It reaches the 14th position (i.e., the worst rank shared with six other physicians) in the order imposed by $R^{*,P}$, while being ranked in the bottom four according to all remaining measures. When compared to the results for group $G1$, significant changes in the outcomes attained for multiple scenarios considered jointly can be noted for MD12. For $G1$, MD12 was ranked outside the top 10 according to all measures. When considering all groups jointly, this happens for only two measures (see $d^{*,P}$ and E_*). Moreover, for some indicators, MD12 is ranked at the very top (see d_*^P and R_*^P). Such differences are implied by the relatively poor performance of MD12 for $G1$ and its favorable evaluation for other complaint groups.

Table A12

Ranks attained by physicians in the orders imposed by different measures derived from robustness analysis for complaint groups G_1 , G_2 , and G_3

MD	d_*^P	$d^{*,P}$	$E^{*,P}$	E_*^P	R_*^P	$R^{*,P}$	$ \tilde{\chi}^{N,P} $	$ \tilde{\chi}^{N,P} $
1	1	20	16	19	1	14	18	15
2	20	6	20	17	20	14	19	19
3	19	19	17	20	19	14	19	19
4	10	18	2	12	9	14	13	11
5	1	13	13	5	1	4	2	2
6	1	9	6	1	1	1	5	2
7	15	5	14	9	18	6	10	11
8	9	7	7	14	9	4	13	11
9	16	10	11	6	17	6	13	11
10	12	17	18	18	12	14	11	17
11	1	12	12	2	1	14	8	7
12	1	14	8	15	1	6	2	7
13	11	1	9	3	9	1	6	2
14	17	11	16	16	15	12	11	16
15	1	2	4	8	1	6	6	6
16	1	16	1	10	1	14	1	2
17	18	8	15	13	12	6	13	10
18	14	4	19	7	15	12	13	18
19	13	3	10	11	12	3	9	7
20	1	15	3	4	1	6	2	1

Publication [P5]

A. Labijak-Kowalska and M. Kadziński. Exact and stochastic methods for robustness analysis in the context of imprecise data envelopment analysis. *Operational Research*, 23(1):22, Mar 2023, DOI: 10.1007/s12351-023-00755-z.

Number of citations⁵:

- according to Google Scholar: 1

⁵as on May 30, 2023



Exact and stochastic methods for robustness analysis in the context of Imprecise Data Envelopment Analysis

Anna Labijak-Kowalska¹ · Miłosz Kadziński¹

Received: 7 September 2021 / Revised: 6 October 2022 / Accepted: 30 October 2022
© The Author(s) 2023

Abstract

We consider the problem of measuring the efficiency of decision-making units with a ratio-based model. In this perspective, we introduce a framework for robustness analysis that admits both interval and ordinal performances on inputs and outputs. The proposed methodology exploits the uncertainty related to the imprecise data and all feasible input/output weight vectors delimited through linear constraints. We offer methods for verifying the robustness of three types of outcomes: efficiency scores, efficiency preference relations, and efficiency ranks. On the one hand, we formulate mathematical programming models to compute the extreme, necessary, and possible results. On the other hand, we incorporate the stochastic analysis driven by the Monte Carlo simulations to derive the probability distribution of different outcomes. The framework is implemented in R and made available on open-source software. Its use is illustrated in two case studies concerning Chinese ports or industrial robots.

Keywords Data Envelopment Analysis · Imprecise performances · Robustness analysis · Monte Carlo simulation · Open-source software

1 Introduction

Data Envelopment Analysis (DEA) measures the relative efficiency of Decision Making Units (DMUs) (Cooper et al. 2014). The standard Charnes-Cooper-Rhodes (CCR) model used in DEA generalizes the single output/input productivity measure (Farrell 1957) by transforming the characterization of each DMU in terms of

✉ Miłosz Kadziński
milosz.kadzinski@cs.put.poznan.pl

Anna Labijak-Kowalska
anna.labijak@cs.put.poznan.pl

¹ Faculty of Computing and Telecommunications, Poznań University of Technology, Piotrowo 2, 60-965 Poznan, Poland

multiple desired outputs and multiple input factors (Charnes et al. 1978). Specifically, the efficiency is quantified as a ratio between a single virtual output and a single virtual input (Salo and Punkka 2011). When evaluating the efficiency of each DMU, the weights involved in the definition of efficiency measure are selected to identify the most advantageous scenario. This means that an efficiency score of a given DMU is maximized subject to both the constraint that all DMUs can have scores lesser or equal to the unity and the feasibility of input/output weights. As a result, DEA generates relative efficiency measures, which depend on the set of analyzed DMUs, leading to the identification of the so-called efficient frontier (Charnes et al. 1994). The units that lie on the frontier attain the score of one. In contrast, the units with a score lesser than one are below the efficient frontier, hence being classified as inefficient.

The main advantages of DEA derive from its following features (Charnes et al. 1994). First, DEA conducts a detailed analysis of performance measures for each DMU instead of focusing on the population averages. This allows for understanding the status of efficiency for individual observations. Moreover, in the case of inefficiency, one could identify its sources and point out the desired modifications of inputs and/or outputs for projecting the DMU onto the efficient frontier (see Aparicio et al. 2007; Chen and Wang 2020; Wu et al. 2018). Second, DEA does not involve any assumption about the functional form, hence not relating the independent and dependent variables (i.e., inputs and outputs) in any specific way. In turn, it evaluates each DMU relative to other DMUs, while not requiring any prior specification of weights. Finally, a great advantage of DEA lies in its simplicity and generality. It captures the efficiency in utilizing the inputs to produce the outputs, all expressed in various units, with a single, easily interpretable performance measure.

For the last forty years, many extensions of DEA have been proposed (see Cook and Seiford 2009; Emrouznejad and Yang 2018). The traditional DEA models assumed that the consumed inputs and produced outputs could be precisely expressed with numerical values on a ratio scale. However, in many real-world problems, this is not possible for a few reasons (see Aparicio et al. 2019; Cooper et al. 1999; Shokouhi et al. 2010). These reasons include inexact specification of inputs and outputs, the uncertainty of data used to compute the consumed inputs or desired outputs, subjectivity involved in this process, and high costs in terms of time or financial resources needed for conducting the accurate measurements (Corrente et al. 2017). As a result, the measurements of inputs and outputs often remain imperfect. This, in turn, requires methodological developments that could handle such uncertain or inaccurate evaluations.

In the context of DEA, two types of imperfect inputs and outputs received particular attention (Liu et al. 2013). On the one hand, the basic idea to capture the uncertainty is using an imprecise evaluation in terms of the interval of possible values. On the other hand, ordinal assessments can be considered. The latter is helpful if only qualitative information is available, some binary features are involved in the analysis, or it is possible to obtain the ranking of units in terms of some input or output instead of precise quantitative measurements.

To handle the imprecision of inputs and outputs, Cooper et al. (1999) proposed Imprecise DEA (IDEA), where precise performance values were replaced with

intervals. This methodology has been further revised and enriched in different ways. For example, Kim et al. (1999) accounted for strong and weak ordinal relations as well as ratio interval data. Furthermore, Despotis and Smirlis (2002) dealt with transforming the interval performances into precise ones, incorporating them into a standard DEA model to optimize the computational performance of the problem. Moreover, Zhu (2003) developed a linear programming model handling strong ordinal inputs and outputs. Also, Ebrahimi and Khalili (2018) proposed the models—incorporating preference information—that find the most preferred DMU and rank other efficient DMUs. The DEA models handling imprecise data have been successfully used in the telecommunication sector (Cooper et al. 2001), machinery industry (Kao and Liu 2005), wheat farming (Hadi-Vencheh and Matin 2011), port efficiency assessment (Zahran et al. 2020), and healthcare (see Azadi and Saen 2013; Karsak and Karadayi 2017).

In the traditional DEA and IDEA methods, only the most favorable input/output weight vector is considered when evaluating each DMU's performance. This may be criticized for a few reasons. First, choosing the individual weight vector for each DMU makes the comparison of efficiencies questionable due to the non-uniqueness of the most advantageous weight vectors and lack of a common basis to analyze the attained scores (Lahdelma and Salminen 2006). Second, such an analysis is focused on a minimal set of scenarios while ignoring other feasible weight vectors that could provide helpful information on the variety of efficiency scores (Salo and Punkka 2011). Third, the efficient frontier, which forms the basis for evaluating the DMUs, requires prior assumptions of the return-to-scale. Besides, it strongly depends on the set of considered DMUs (see Zhu 1996; Seiford and Zhu 1998). Fourth, using a single efficiency measure that divides the DMUs into efficient and inefficient ones offers too limited capabilities for discriminating between the units (see Adler et al. 2002; Hosseinzadeh Lotfi et al. 2013). All these drawbacks motivated the development of robustness analysis methods, which quantify the stability of efficiency results for different feasible weight vectors. Given imprecise inputs and outputs, the need to include uncertainties when working out the results is even more evident. The robust conclusions should be valid in all or most scenarios (see Kadziński and Tervonen 2013; Liang et al. 2020), with a scenario being equivalent to a set of possible values for data of the problem and the efficiency model parameters.

Some essential methodological advancements oriented toward robustness analysis for IDEA have been proposed over the last two decades. In particular, Despotis and Smirlis (2002) derived the optimistic and pessimistic efficiency scores for each DMU. Both are computed with the most favorable weight vectors for a given unit while assuming the most and the least advantageous scenarios for the inputs and outputs. Based on these results, the units can be divided into three groups: efficient in the most pessimistic scenario, inefficient even in the most optimistic scenario, and an intermediate class including DMUs with unitary optimistic efficiency and pessimistic efficiency lesser than one. This classification was further analyzed in Jahanshahloo et al. (2004) to consider the “radius of stability”. For each DMU, it is defined with a pair of values, α and β , indicating, respectively, a decrease of the upper bounds of input and output intervals and an increase of the respective lower bounds for which the efficiency class remains unchanged. Furthermore, Kao

(2006) proposed mathematical models for computing the optimistic and pessimistic efficiency scores in the presence of both interval and ordinal inputs and outputs. A similar aim of deriving an efficiency interval for each DMU—though in different settings—was considered in Ebrahimi and Toloo (2020) and Park (2007). Also, in this context, Ebrahimi et al. (2021) and Toloo et al. (2021) accounted for the dual-role factors, which can be interpreted as input and output at the same time. In turn, Haghighat and Khorram (2005) proposed non-linear models for deriving the maximal and minimal numbers of efficient units when the input and output performances are given as intervals. The Monte Carlo simulation was incorporated into the stochastic DEA to derive the distribution of efficiency scores in the setting where inputs and outputs were expressed as intervals formed by values gathered in different years (Kao and Liu 2009). Dehnokhalaji et al. (2022) proposed a robust optimization framework for performance measurement and cross-efficiency inspired ranking of DMUs. An additive value DEA model was considered in Gouveia et al. (2013) to construct the efficiency intervals and find the maximal percentage tolerance by which one could deteriorate the inputs or outputs of a given DMU so that it remains efficient. Finally, Azizi et al. (2015) proposed a slack-based method to find the optimistic and pessimistic efficiency intervals for DMUs for DEA involving imprecise data. Specifically, two classifications of DMUs into efficient and inefficient units were proposed considering the optimistic and pessimistic settings. In addition, the procedures for obtaining an overall interval score as well as constructing a complete ranking of DMUs were introduced.

The most important contribution of this paper consists of proposing a rich framework for robustness analysis in the context of imprecise inputs and outputs. As opposed to the existing approaches that extend IDEA, our methodology considers uncertainty related to the interval or ordinal data and all feasible weight vectors simultaneously. In particular, we propose tools for analyzing the robustness of three types of outcomes: efficiency scores, efficiency preference relations, and efficiency ranks.

On the one hand, we derive extreme, robust results using dedicated mathematical programming models exploiting all scenarios involving imprecise input/output data and feasible weight vectors. We show how to compute the extreme efficiency scores and ranks and verify the truth or falsity of the necessary and possible efficiency preference relations (Kadziński et al. 2017). The efficiency bounds and ranking intervals reveal the pessimistic and optimistic performance of each unit (Salo and Punkka 2011). In turn, the two relations focus on the pairwise comparisons that need to be validated for all or at least one feasible scenario (Kadziński et al. 2017).

On the other hand, we implement the stochastic analysis to derive the distribution of different measures and results (Lahdelma and Salminen 2006). We employ the Monte Carlo simulations to analyze a sufficiently large and representative set of feasible weight vectors and input/output performances consistent with the imprecise information. For this purpose, we apply a suitably adjusted Hit-And-Run algorithm (see Ciomek and Kadziński 2021; Tervonen et al. 2013). The outcomes are quantified through Efficiency Acceptability Interval Indices, Efficiency Rank Acceptability Indices, and Pairwise Efficiency Outranking Indices (see Lahdelma and Salminen 2006; Kadziński et al. 2017). The stochastic indices capture the shares of feasible

scenarios that guarantee a given score or rank to a particular DMU or confirm that one DMU is at least as good as the other. Also, we estimate the expected efficiency scores and ranks for all DMUs. These measures can be the basis for constructing a complete ranking of DMUs based on the robust outcomes derived from analyzing feasible weights, inputs, and outputs. From the methodological perspective, the proposed methodology can be seen as an extension and adjustment of an integrated framework for robustness analysis proposed in Kadziński et al. (2017) to the case of imprecise (interval or ordinal) evaluations.

We also present open-source software that implements the proposed framework for robustness analysis. The software consists of modules available on the *diviz* platform (Meyer and Bigaret 2012). These modules accept the specification of linear constraints concerning the weights related to inputs and outputs. Moreover, they have been designed to admit their combination into complex algorithmic workflows. The latter can be employed to share the methodological developments and results of case studies among users.

Finally, we illustrate the use of both the framework for robustness analysis and software in real-world studies concerning efficiency analysis of Chinese ports (Jiang et al. 2021) and industrial robots (Saen 2006). The units are described in terms of precise, interval, and ordinal factors. These examples demonstrate the practical usefulness of robust results concerning scores, ranks, and pairwise preference relations. Also, we emphasize the complementarity of exact and stochastic results. Moreover, we demonstrate that both the space of feasible weight vectors as well as imprecise input and output performances influence the robustness of attained efficiency results.

The remainder of the paper is organized in the following way. In Sect. 2, we discuss the proposed methods for robustness analysis within the scope of Imprecise Data Envelopment Analysis. In Sect. 3, we present the algorithmic modules implementing the proposed methodological framework on the *diviz* platform. Section 4 is devoted to an illustrative case study concerning the efficiency analysis of Chinese ports. The results of the study on industrial robots are reported in the e-Appendix (supplementary material available online). Section 5 concludes the paper and outlines avenues for future work.

2 Robustness analysis for Imprecise Data Envelopment Analysis

2.1 Notation and basic concepts

The following notation is used in the paper:

- $\mathcal{D} = \{DMU_1, \dots, DMU_K\}$ —a set of considered DMUs, where K is the number of DMUs ($K = |\mathcal{D}|$);
- x_m — m -th input, $m \in \{1, \dots, M\}$;
- y_n — n -th output, $n \in \{1, \dots, N\}$;
- PI , II and OI —subsets of precise, interval, and ordinal inputs, respectively;
- PO , IO and OO —subsets of precise, interval, and ordinal outputs, respectively;
- x_{mo} —the value of m -th input consumed by $DMU_o \in \mathcal{D}$, $m \in PI \cup OI$;

- y_{no} —the value of n -th output produced by $DMU_o \in \mathcal{D}$, $n \in PO \cup OO$;
- $[x_{mo*}, x_{mo}^*]$ —an interval value of m -th input of DMU_o , $m \in II$;
- $[y_{no*}, y_{no}^*]$ —an interval value of n -th output of DMU_o , $n \in IO$;
- X_{mo} —the value of $v_m \cdot x_{mo}$ for ordinal inputs, $m \in OI$;
- Y_{no} —the value of $u_n \cdot y_{no}$ for ordinal outputs, $n \in OO$;
- $v = \{v_1, \dots, v_M\}$ —a vector of input weights;
- $u = \{u_1, \dots, u_N\}$ —a vector of output weights;
- η, χ —values representing the minimal ratios between the successive values of ordinal inputs and ordinal outputs, $\eta, \chi > 1$ (in this paper, we set $\eta = \chi = 1.1$);
- $S_v = \{v = (v_1, \dots, v_M)^T \neq 0 \mid v \geq 0, A_v v \leq 0\}$ and $S_u = \{u = (u_1, \dots, u_N)^T \neq 0 \mid u \geq 0, A_u u \leq 0\}$ —spaces of feasible input and output weights, respectively; A_v and A_u are matrices of coefficients involved in the linear constraints on weights derived from the user's preferences.

To illustrate the notation, let us refer to an example presented in Table 1, which is derived from Despotis and Smirlis (2002). The set of DMUs is composed of five units, $\mathcal{D} = \{D_1, D_2, D_3, D_4, D_5\}$. They consume two inputs—one precise ($PI = \{i_1\}$) and the other interval ($II = \{i_2\}$), and produce two outputs—one precise ($PO = \{o_1\}$) and the other ordinal ($OO = \{o_2\}$). The weights associated with the inputs are denoted by v_1 and v_2 , and the respective weights for the outputs are u_1 and u_2 . When it comes to unit D_1 , its precise input is $x_{11} = 100$ and the interval input is $[x_{21*}, x_{21}^*] = [0.6, 0.7]$. The respective outputs are $y_{11} = 2000$ and $y_{21} = 4$. The latter will be represented in the following mathematical models as $Y_{21} = y_{21} \cdot u_2$, and the following order $Y_{24} < Y_{22} < Y_{25} < Y_{21} < Y_{23}$ will be maintained.

In what follows, we discuss the methods for robustness analysis in the context of Imprecise DEA. They can be divided into two subgroups. One of them is devoted to the exact analysis using linear programming techniques. In contrast, the other aims to estimate some stochastic acceptability indices through the Monte Carlo simulations. The analysis is conducted given all feasible efficiency scenarios, where each scenario corresponds to a specific, admissible realization of both weights and performances on inputs and outputs.

Table 1 Example set of Decision Making Units involving imprecise data

DMU_o	i_1 (precise)	i_2 (interval)	o_1 (precise)	o_2 (ordinal)
D_1	100	[0.6, 0.7]	2000	4
D_2	150	[0.8, 0.9]	1000	2
D_3	150	[1.0, 1.0]	1200	5
D_4	200	[0.7, 0.8]	900	1
D_5	200	[1.0, 1.0]	600	3

2.2 Exact robustness analysis with linear programming

In this section, we discuss how to derive exact robust outcomes using mathematical programming. These results capture the extreme cases observed for all feasible efficiency scenarios $(\mathbf{u}, \mathbf{v}, \mathbf{x}, \mathbf{y})$ defined by the sets of admissible weights as well as values of inputs and outputs. They concern the following three perspectives: scores, ranks, and pairwise preference relations. We refer to the concepts of extreme scores and ranks and the necessary and possible preference relations that have been introduced in the literature. However, the models for their computation that are presented in this section are original and specifically adjusted to the context of IDEA.

To conduct a robustness analysis given interval inputs and outputs, we need to consider the most and the least advantageous (i.e., optimistic and pessimistic) scenarios for each DMU. On the one hand, the optimistic scenario for DMU_o is realized by assuming that its inputs are the least possible and its outputs are the greatest admissible by the specified intervals. In contrast, for the remaining units, both the inputs and outputs are the least advantageous, i.e.:

$$x_{mk} = \begin{cases} x_{mk*}, & \text{if } m = o, \\ x_{mk}^*, & \text{otherwise,} \end{cases} \quad (1)$$

$$y_{nk} = \begin{cases} y_{nk}^*, & \text{if } n = o, \\ y_{nk*}, & \text{otherwise.} \end{cases} \quad (2)$$

On the other hand, the pessimistic scenario for DMU_o is realized by assuming that its imprecise inputs and outputs are replaced with the least favorable values. For the remaining DMUs, the minimal inputs and the maximal outputs are considered, i.e.:

$$x_{mk} = \begin{cases} x_{mk}^*, & \text{if } m = o, \\ x_{mk*}, & \text{otherwise,} \end{cases} \quad (3)$$

$$y_{nk} = \begin{cases} y_{nk*}, & \text{if } n = o, \\ y_{nk}^*, & \text{otherwise.} \end{cases} \quad (4)$$

When the dataset involves the ordinal factors, the products $v_m \cdot x_{mk}$ or $u_n \cdot y_{nk}$ are replaced by one variable, respectively, X_{mk} or Y_{nk} . Additionally, the constraints respecting the character of ordinal evaluations need to be included in the model. In particular, the constraints imposing a strong ordinal relation should not take an additive form, e.g., $X_2 \geq X_1 + \epsilon$, where ϵ is a small positive constant. In turn, as the original ordinal evaluations x_{mk} and y_{nk} are transformed into variables X_{mk} or Y_{nk} involving multiplication by a common weight (v_m and u_n), the ratios of subsequent X_{mk} or Y_{nk} values needs to be greater than one, i.e.:

$$\begin{cases} \chi \cdot Y_{ni} \leq Y_{nj}, & (i, j) \in \{(i, j) : y_{ni} \leq y_{nj}\}, n \in OO, \\ \eta \cdot X_{mi} \leq X_{mj}, & (i, j) \in \{(i, j) : x_{mi} \leq x_{mj}\}, m \in OI, \end{cases} \quad (5)$$

where $\chi, \eta > 1$ (Zhu 2003).

Following (Zhu 2003), we consider the efficiency of DMU_o defined as a ratio of a single virtual output to a single virtual input:

$$E_o = \frac{\sum_{n \in PO} u_n y_{no} + \sum_{n \in IO} u_n y_{no} + \sum_{n \in OO} Y_{no}}{\sum_{m \in PI} v_m x_{mo} + \sum_{m \in II} v_m x_{mo} + \sum_{m \in OI} X_{mo}} \quad (6)$$

where $y_{no} \in [y_{no*}, y_{no}^*]$ for $n \in IO$ and $x_{mo} \in [x_{mo*}, x_{mo}^*]$ for $m \in II$. The virtual output and input aggregate multiple outputs or inputs while ensuring that each relevant factor contributes to an overall measure of efficiency. Please note that the contributions from these factors are dimensionless. This is due to multiplying the precise and imprecise performances by the weights and using dedicated components for the ordinal factors. Still, their major role is to maintain the desired relationships between the efficiencies of various units implied by their input and output values. In fact, the above expression ensures that E_o does not deteriorate if one (i) increases the output values or decreases the input values in DMU_o or (ii) decreases the output values or increases the input values in other DMUs. At the same time, this representation eliminates the scale transformations (Zhu 2003), reducing the computational burden in applications.

2.2.1 Extreme efficiency scores

When it comes to the efficiencies, for each DMU_o , we determine the maximal E_o^* and minimal E_{o*} scores that it can attain for at least one feasible scenario (see Despotis and Smirlis 2002; Kadziński et al. 2017; Kao 2006). To find the greatest (optimistic) efficiency score for DMU_o , the following Linear Programming (LP) model needs to be solved:

$$\begin{aligned} & \text{Maximize: } E_o^* = \sum_{n \in PO} u_n y_{no} + \sum_{n \in IO} u_n y_{no}^* + \sum_{n \in OO} Y_{no} \\ \text{s.t. } [E^* - C1] & \quad \sum_{m \in PI} v_m x_{mo} + \sum_{m \in II} v_m x_{mo*} + \sum_{m \in OI} X_{mo} = 1, \\ [E^* - C2] & \quad \sum_{n \in PO} u_n y_{no} + \sum_{n \in IO} u_n y_{no}^* + \sum_{n \in OO} Y_{no} \leq 1, \\ [E^* - C3] & \quad \sum_{n \in PO} u_n y_{nk} + \sum_{n \in IO} u_n y_{nk*} + \sum_{n \in OO} Y_{nk} \leq \sum_{m \in PI} v_m x_{mk} + \sum_{m \in II} v_m x_{mk}^* + \sum_{m \in OI} X_{mk}, \quad k = 1, \dots, K; k \neq o, \\ [E^* - C4] & \quad \chi Y_{ni} \leq Y_{nj}, \quad (i, j) \in \{(i, j) : y_{ni} \leq y_{nj}\}, n \in OO, \\ [E^* - C5] & \quad \eta X_{mi} \leq X_{mj}, \quad (i, j) \in \{(i, j) : x_{mi} \leq x_{mj}\}, m \in OI, \\ [E^* - C6] & \quad (v, u) \in (S_v, S_u). \end{aligned} \quad (7)$$

Model (7) is equivalent to the classical CCR model for DEA with imprecise data. It finds the most favorable weight vector for DMU_o in its best input/output scenario and the worst possible scenarios for the remaining DMUs. The space of variables is composed of the following weights: v_m for $n \in PI \cup II$, X_{mk} for $m \in OI$ and $DMU_k \in \mathcal{D}$, u_n for $n \in PO \cup IO$ and Y_{nk} for $n \in OO$ and $DMU_k \in \mathcal{D}$. It is constrained so that the virtual input of DMU_o is equal to one ($[E^* - C1]$), the efficiency scores for all DMUs are not greater than one ($[E^* - C2]$ and $[E^* - C3]$), the monotonicity relations derived from the analysis of ordinal inputs and outputs are preserved ($[E^* - C4]$ and $[E^* - C5]$), and the constraints on the admissible values of input and output weights are satisfied ($[E^* - C6]$). The last three constraints are present in all

following LP models. The optimal value of E_o^* is between zero and one. The DMUs with optimal $E_o^* = 1$ are considered as efficient.

The minimal (pessimistic) efficiency score for DMU_o using the CCR model with imprecise information can be derived by solving the following Mixed-Integer Linear Programming (MILP) model:

$$\begin{aligned}
 & \text{Minimize: } E_{o^*} = \sum_{n \in PO} u_n y_{no} + \sum_{n \in IO} u_n y_{nos} + \sum_{n \in OO} Y_{no} \\
 \text{s.t. } [E_* - C1] & \quad \sum_{m \in PI} v_m x_{mo} + \sum_{m \in II} v_m x_{m^*o} + \sum_{m \in OI} X_{mo} = 1, \\
 [E_* - C2] & \quad \sum_{n \in PO} u_n y_{no} + \sum_{n \in IO} u_n y_{nos} + \sum_{n \in OO} Y_{no} \geq 1 - C(1 - b_o), \\
 [E_* - C3] & \quad \sum_{n \in PO} u_n y_{nk} + \sum_{n \in IO} u_n y_{nk^*} + \sum_{n \in OO} Y_{nk} \geq \sum_{m \in PI} v_m x_{mk} + \sum_{m \in II} v_m x_{mk^*} + \sum_{m \in OI} X_{mk} - C(1 - b_k), \quad k = 1, \dots, K; k \neq o, \\
 [E_* - C4] & \quad \sum_{k=1}^K b_k \geq 1, \\
 [E_* - C5] & \quad b_k \in \{0, 1\}, \quad k = 1, \dots, K, \\
 [E_* - C6] & \quad \chi Y_{ni} \leq Y_{nj}, \quad (i, j) \in \{(i, j) : y_{ni} \leq y_{nj}\}, n \in OO, \\
 [E_* - C7] & \quad \eta X_{mi} \leq X_{mj}, \quad (i, j) \in \{(i, j) : x_{mi} \leq x_{mj}\}, m \in OI, \\
 [E_* - C8] & \quad (v, u) \in (S_v, S_u).
 \end{aligned} \tag{8}$$

The above model allows for finding the least favorable weight vector for DMU_o in terms of its efficiency while considering the worst possible scenario for DMU_o and the best admissible scenarios for the remaining units $DMU_k, k = 1, \dots, K$ and $k \neq o$. Under these conditions, we constrain the space of feasible solutions by imposing—without loss of generality—that the virtual input of DMU_o equals one ($[E_* - C1]$), assuming that at least one unit is efficient (its efficiency score must be greater than or equal to one; $[E_* - C2]$ – $[E_* - C5]$), preserving the ordinal factors’ monotonicity ($[E_* - C6]$ – $[E_* - C7]$), and satisfying the pre-defined constraints on the admissible values of input and output weights ($[E_* - C8]$). Apart from the weights already considered in model (7), we include the binary variables $b_k \in \{0, 1\}, k = 1, \dots, K$. The optimal value of E_{o^*} is between zero and one. Overall, $[E_{o^*}, E_o^*]$ can be deemed as an efficiency interval (Salo and Punkka 2011).

Note that C is a large positive constant. Irrespective of which DMU_o is considered, it is sufficient that $C > \max_{DMU_l, DMU_k \in \mathcal{D}} \{ \max\{ \max_{m \in PI} \{x_{mk}/x_{ml}\}, \max_{m \in II} \{x_{mk^*}/x_{ml^*}\}, \text{ if } OI \neq \emptyset : \eta^K \} \}$. It is so because to minimize E_{o^*} , the solver also minimizes E_k for $k = 1, \dots, K$. Since constraint $[E_* - C1]$ imposes $\sum_{m \in PI} v_m x_{mo} + \sum_{m \in II} v_m x_{m^*o} + \sum_{m \in OI} X_{mo} = 1$, then for $k = 1, \dots, K, C$ is greater than $\sum_{m \in PI} v_m x_{mk} + \sum_{m \in II} v_m x_{mk^*} + \sum_{m \in OI} X_{mk}$. Consequently, when binary variable b_k equals 0 for $k = 1, \dots, K$, constraint $[E_* - C2]$ (when $k = o$) or constraint $[E_* - C3]$ (when $k = 1, \dots, K, k \neq o$) is satisfied for all values of the variables. However, constraint $[E_* - C4]$ imposes that at least one b_k for $k = 1, \dots, K$ is equal to one. Then, the respective efficiency E_k is greater or equal to one since the virtual output of DMU_k is greater or equal to its virtual input.

Illustrative example In Table 2, we present the extreme efficiencies derived for five units contained in the illustrative example introduced in Sect. 2.1. They reveal that two units (D_1 and D_3) are efficient, attaining the maximal efficiency score equal to one. The

Table 2 Exact robust results derived with mathematical programming for the illustrative example

DMU_o	Extreme scores		Extreme ranks		Robust relations				
	E_o^*	E_{o*}	R_o^*	R_{o*}	D_1	D_2	D_3	D_4	D_5
D_1	1.000	0.013	1	2	N	N	P	N	N
D_2	0.723	0.011	2	5		N	P	P	P
D_3	1.000	0.367	1	4	P	P	N	P	N
D_4	0.751	0.010	2	5		P	P	N	P
D_5	0.637	0.012	3	5		P		P	N

efficiency intervals are relatively wide and span over the range of over 0.6 for all units. For example, the minimal efficiencies of D_1 and D_3 are, respectively, 0.013 and 0.367.

2.2.2 Extreme efficiency ranks

As far as efficiency ranks are concerned, we determine the best R_o^* and the worst R_{o*} ranks that are attained by DMU_o for at least one feasible scenario (see Kadziński et al. 2012, 2017; Salo and Punkka 2011). Given a fixed input/output weight vector and precise feasible performances for DMU_o , it attains k -th rank if exactly $k - 1$ other units attain higher efficiency scores than DMU_o . To find the minimal (i.e., the best) efficiency rank for DMU_o , the following MILP model needs to be solved:

$$\begin{aligned}
 &\text{Minimize:} && R_o^* = 1 + \sum_{k=1, k \neq o}^K b_k \\
 \text{s.t. } &[R^* - C1] && \sum_{n \in PO} u_n y_{no} + \sum_{n \in IO} u_n y_{no}^* + \sum_{n \in OO} Y_{no} = 1, \\
 &[R^* - C2] && \sum_{m \in PI} v_m x_{mo} + \sum_{m \in II} v_m x_{mo}^* + \sum_{m \in OI} X_{mo} = 1, \\
 &[R^* - C3] && \sum_{n \in PO} u_n y_{nk} + \sum_{n \in IO} u_n y_{nk}^* + \sum_{n \in OO} Y_{nk} \leq \sum_{m \in PI} v_m x_{mk} + \sum_{m \in II} v_m x_{mk}^* + \sum_{m \in OI} X_{mk} + Cb_k, && k = 1, \dots, K; k \neq o, \\
 &[R^* - C4] && b_k \in \{0, 1\}, && k = 1, \dots, K; k \neq o, \\
 &[R^* - C5] && \chi Y_{ni} \leq Y_{nj}, && (i, j) \in \{(i, j) : y_{ni} \leq y_{nj}\}, n \in OO, \\
 &[R^* - C6] && \eta X_{mi} \leq X_{mj}, && (i, j) \in \{(i, j) : x_{mi} \leq x_{mj}\}, m \in OI, \\
 &[R^* - C7] && (v, u) \in (S_v, S_u).
 \end{aligned} \tag{9}$$

The above model sets the efficiency score of DMU_o in its most optimistic realization equal to one ($[R^* - C1]$ – $[R^* - C2]$). For the remaining units, we assume their pessimistic realizations ($[R^* - C3]$) and minimize the number of DMUs with efficiency scores greater than for DMU_o . This is attained by introducing the binary variables b_k for each DMU_k , $k = 1, \dots, K$, and $k \neq o$ ($[R^* - C5]$). When the efficiency score of DMU_k cannot be lower than or equal to one, then b_k is set to one, and the respective constraint $[R^* - C4]$ is always satisfied for all possible variable values. This is implied by the use of a large positive constant C . Analogously to the reasoning for model (8), irrespective of which DMU_o is considered, it is sufficient that $C > \max_{DMU_1, DMU_k \in D} \{ \max\{ \max_{n \in PO} \{y_{nk}/y_{nl}\}, \max_{n \in IO} \{y_{nk}^*/y_{nl}^*\} \}, \text{ if } OO \neq \emptyset : \chi^K \}$. In turn, if the efficiency score of DMU_o is greater or equal to the efficiency of DMU_k , b_k is set to zero. Thus, by minimizing the

sum of $b_k, k = 1, \dots, K$, and $k \neq o$, we can obtain the best possible rank of DMU_o . The optimal value of R_o^* is between one and K .

The worst (i.e., the maximal) possible rank of DMU_o can be computed with the following MILP model:

$$\begin{aligned}
 &\text{Maximize:} && R_{os} = 1 + \sum_{k=1, k \neq o}^K b_k \\
 \text{s.t. } &[R_* - C1] && \sum_{n \in PO} u_n y_{no} + \sum_{n \in IO} u_n y_{no*} + \sum_{n \in OO} Y_{no} = 1, \\
 &[R_* - C2] && \sum_{m \in PI} v_m x_{mo} + \sum_{m \in II} v_m x_{mo}^* + \sum_{m \in OI} X_{mo} = 1, \\
 &[R_* - C3] && \sum_{m \in PI} v_m x_{mk} + \sum_{m \in II} v_m x_{mk}^* + \sum_{m \in OI} X_{mk} \leq \sum_{n \in PO} u_n y_{nk} + \sum_{n \in IO} u_n y_{nk}^* + \sum_{n \in OO} Y_{nk} + C(1 - b_k), && k = 1, \dots, K; k \neq o, \\
 &[R_* - C4] && b_k \in \{0, 1\}, && k = 1, \dots, K; k \neq o, \\
 &[R_* - C5] && \chi Y_{ni} \leq Y_{nj}, && (i, j) \in \{(i, j) : y_{ni} \leq y_{nj}\}, n \in OO, \\
 &[R_* - C6] && \eta X_{mi} \leq X_{mj}, && (i, j) \in \{(i, j) : x_{mi} \leq x_{mj}\}, m \in OI, \\
 &[R_* - C7] && (v, u) \in (S_v, S_u).
 \end{aligned} \tag{10}$$

The above model maximizes the number of DMUs with efficiency scores greater or equal to DMU_o . Again, we assume that the efficiency of DMU_o is equal to one ($[R_* - C1]$ – $[R_* - C2]$). However, at this time, we consider the pessimistic realization of DMU_o . Then, we introduce the constraints imposing that the efficiencies of the remaining DMUs in their optimistic realizations are not lower than one ($[R_* - C3]$). The component $C \cdot (1 - b_k)$ included in the respective constraint implies that the latter can be violated. If binary variable b_k ($[R_* - C4]$) is equal to one, constraint $[R_* - C3]$ holds, whereas for $b_k = 0$ —it is satisfied for any variables’ values. Note that C should be set similarly as for model (8). When maximizing the sum of $b_k, k = 1, \dots, K$, and $k \neq o$, we minimize the number of DMUs for which constraint $[R_* - C3]$ is violated. Thus, the sum of b_k increased by one corresponds to the worst possible rank of DMU_o . The optimal value of R_{o*} is between one and K .

Illustrative example The extreme ranks for the illustrative example introduced in Sect. 2.1 are presented in Table 2. The efficient units attain the first rank in the best case. Although the minimal efficiency of D_1 is worse than for D_3 , in the worst case its rank can drop only to the second position ($R_{1*} = 2$), whereas D_3 can be ranked even fourth ($R_{3*} = 4$) in the most pessimistic scenario. The inefficient units can be ranked second (D_2 and D_4) or third (D_5) in the best case, while all are ranked at the bottom in the least advantageous scenario.

2.2.3 Necessary and possible efficiency preference relations

When it comes to the stability of comparisons observed for pairs of DMUs given all feasible scenarios, we consider the necessary (\succsim_E^N) and possible (\succsim_E^P) efficiency preference relations (see Greco et al. 2008; Kadziński et al. 2017). They are defined in the following way:

- DMU_o is necessarily preferred to DMU_l ($DMU_o \succsim_E^N DMU_l$) if DMU_o attains at least as good efficiency as DMU_l for all feasible scenarios defined by the sets of admissible weights, as well as values of inputs and outputs, or, equivalently, if

for all feasible weight vectors the efficiency of DMU_o in its pessimistic realization is not worse than the efficiency of DMU_l in its optimistic realization;

- DMU_o is possibly preferred to DMU_l ($DMU_o \succeq_E^P DMU_l$) if DMU_o attains at least as good efficiency as DMU_l for at least one feasible scenario defined by the sets of admissible weights, as well as values of inputs and outputs, or, equivalently, if for at least one feasible weight vector the efficiency of DMU_o in its optimistic realization is not worse than the efficiency of DMU_l in its pessimistic relations.

To verify the truth of the necessary efficiency preference relation $DMU_o \succeq_E^N DMU_l$ for pair (DMU_o, DMU_l) , we need to solve the following LP model:

$$\begin{aligned}
 \text{Minimize: } E_{o^*} &= \sum_{n \in PO} u_n y_{no} + \sum_{n \in IO} u_n y_{no^*} + \sum_{n \in OO} Y_{no} \\
 \text{s.t. } \sum_{m \in PI} v_m x_{mo} + \sum_{m \in II} v_m x_{mo^*} + \sum_{m \in OI} X_{mo} &= 1, \\
 \sum_{n \in PO} u_n y_{nl} + \sum_{n \in IO} u_n y_{nl}^* + \sum_{n \in OO} Y_{nl} &= \sum_{m \in PI} v_m x_{ml} + \sum_{m \in II} v_m x_{ml}^* + \sum_{m \in OI} X_{ml}, \\
 \chi Y_{ni} &\leq Y_{nj}, & (i, j) \in \{(i, j) : y_{ni} \leq y_{nj}\}, n \in OO, \\
 \eta X_{mi} &\leq X_{mj}, & (i, j) \in \{(i, j) : x_{mi} \leq x_{mj}\}, m \in OI, \\
 (v, u) &\in (S_v, S_u).
 \end{aligned} \tag{11}$$

The above model finds the minimal efficiency score of DMU_o in its pessimistic realization while assuming that the efficiency of DMU_l in its optimistic realization is equal to one. If the obtained optimal value of E_{o^*} is greater than or equal to one, then for all weight vectors (\mathbf{u}, \mathbf{v}) , the efficiency of DMU_o is not worse than efficiency of DMU_l , i.e., $DMU_o \succeq_E^N DMU_l$. Otherwise, $\text{not}(DMU_o \succeq_E^N DMU_l)$.

The following LP model allows verifying the truth of the possible efficiency preference relation $DMU_o \succeq_E^P DMU_l$ for pair (DMU_o, DMU_l) :

$$\begin{aligned}
 \text{Maximize: } E_o^* &= \sum_{n \in PO} u_n y_{no} + \sum_{n \in IO} u_n y_{no}^* + \sum_{n \in OO} Y_{no} \\
 \text{s.t. } \sum_{m \in PI} v_m x_{mo} + \sum_{m \in II} v_m x_{mo}^* + \sum_{m \in OI} X_{mo} &= 1, \\
 \sum_{n \in PO} u_n y_{nl} + \sum_{n \in IO} u_n y_{nl}^* + \sum_{n \in OO} Y_{nl} &= \sum_{m \in PI} v_m x_{ml} + \sum_{m \in II} v_m x_{ml}^* + \sum_{m \in OI} X_{ml}, \\
 \chi Y_{ni} &\leq Y_{nj}, & (i, j) \in \{(i, j) : y_{ni} \leq y_{nj}\}, n \in OO, \\
 \eta X_{mi} &\leq X_{mj}, & (i, j) \in \{(i, j) : x_{mi} \leq x_{mj}\}, m \in OI, \\
 (v, u) &\in (S_v, S_u).
 \end{aligned} \tag{12}$$

The above model computes the maximal efficiency of DMU_o in its optimistic realization while assuming that the efficiency of DMU_l in its pessimistic realization is equal to one. If the optimal value of E_o^* is greater than or equal to one, then there exists at least one weight vector (\mathbf{u}, \mathbf{v}) for which the efficiency of DMU_o is not worse than efficiency of DMU_l , i.e., $DMU_o \succeq_E^P DMU_l$. Otherwise, $\text{not}(DMU_o \succeq_E^P DMU_l)$.

Illustrative example The necessary and possible relations for the illustrative example are presented in Table 2. Note that the necessary relation is transitive and implies the truth of the possible relation. Let us observe that unit D_1 is necessarily preferred to the three inefficient units (D_2, D_4 , and D_5), whereas D_3 is robustly at least as good only when compared to D_5 . The efficient units are incomparable in terms of the necessary relation while being possibly preferred over each other. The inefficient units are not preferred over any other unit for all feasible scenarios. However, they

are possibly preferred over each other (e.g., $D_2 \succeq_E^P D_4$ and $D_4 \succeq_E^P D_2$). Moreover, D_2 and D_4 are at least as good as D_3 for at least one feasible scenario, whereas none of the inefficient units attains the score of D_1 for any feasible setting.

2.3 Stochastic analysis with the Monte Carlo simulation

In this section, we discuss how to derive stochastic outcomes using the Monte Carlo simulation. These results capture the share or distribution of feasible efficiency scenarios $(\mathbf{u}, \mathbf{v}, \mathbf{x}, \mathbf{y})$ that confirm a given outcome referring to the attained scores or ranks, or the truth of pairwise preference relation.

To conduct such a stochastic analysis, we need to sample a representative subset of all feasible efficiency scenarios. This requires the assumption about the probability distributions of the joint density function of the feasible input/output weight vectors and the precise performances within the specified interval values on various inputs and outputs (Lahdelma and Salminen 2006). In general, the proposed approach can be used with any arbitrarily selected distribution. However, when the expert does not impose the parameter distribution, we assume the uniform distribution of weights and performances (see Kadziński et al. 2017; Lahdelma and Salminen 2001).

To simulate the feasible efficiency scenarios, we need to derive the weights and performances from the feasible space. For sampling weights from the uniform distribution, we use the Hit-And-Run (HAR) algorithm (Tervonen et al. 2013). Since it requires the space of sampling to be bounded, we perform normalization of possible input/output weights:

$$\sum_{n=1}^N u_n = \sum_{m=1}^M v_m = 1. \quad (13)$$

When it comes to sampling the performances, a dedicated treatment has been designed to deal with the interval and ordinal factors. For the interval inputs and outputs, for each DMU_o , we randomly select the exact values from the intervals $[x_{mo}^*, x_{mo}^*]$ or $[y_{no}^*, y_{no}^*]$ using HAR. Regarding dealing with the ordinal factors, we adopt the SMAA-O approach (Lahdelma et al. 2003). Specifically, we assume that a function simulating some ordinal inputs or outputs is increasing. We assume that the exact values corresponding to the ordinal performances are drawn from the $[0, 1]$ interval without losing generality. Hence we randomly choose a set of K numbers from this range. The obtained values are sorted and considered as a single sample of precise performances of DMUs consistent with the order imposed by the original ordinal performances (e.g., a unit with the worst ordinal output or the best ordinal input is assigned the least precise value).

The samples concerning the weights and the input and output values are put together to simulate the feasible efficiency scenarios. For each of them, we compute the efficiencies for all DMUs. The results obtained for all sampled scenarios are summarized in stochastic acceptability indices concerning scores, ranks, and pairwise relations. Since their values are approximated using the Monte Carlo simulation

rather than computed exactly through analytical methods, we consider the estimations of the true indices in practice. However, with a sufficiently large number of samples, such values can be estimated up to a pre-defined accuracy (Tervonen and Lahdelma 2007).

2.3.1 Distribution of efficiency scores

The Efficiency Acceptability Interval Index $EAI(DMU_o, b_i)$ is defined as the share of feasible scenarios for which the efficiency score of DMU_o is contained in the sub-interval $b_i \subset [0, 1]$, where $i = 1, \dots, B$, and B is the number of efficiency sub-intervals considered in the analysis. By default, the sub-intervals are assumed to be disjoint and to span over the same widths. Note that for each $DMU_o \in \mathcal{D}$, $\sum_{i=1}^B EAI(DMU_o, b_i) = 1$. Moreover, by analyzing the scores obtained by DMU_o , we may compute an expected efficiency (denoted by EE_o) as an average of efficiencies derived for all sampled scenarios. Such efficiency may be used to impose a complete ranking on the set of DMUs (Labijak-Kowalska and Kadziński 2021).

Illustrative example In Table 3, we present the estimates of EAI s computed based on 10,000 samples derived with the Monte Carlo simulation for the illustrative example introduced in Sect. 2.1. We have selected five buckets ($B = 5$), and hence the considered sub-intervals are $[0, 0.2]$, $(0.2, 0.4]$, \dots , $(0.8, 1.0]$. The most probable efficiency ranges for D_1 and D_2 are, respectively, $(0.8, 1.0]$ ($EAI(D_1, (0.8, 1.0]) = 0.958$) and $(0.2, 0.4]$ ($EAI(D_2, (0.2, 0.4]) = 0.716$). On the other extreme, the estimated probability of D_1 attaining an efficiency score lower than 0.2 or D_2 attaining a score greater than 0.6 is zero. However, the analysis of extreme efficiency scores presented in Sect. 2.2.1 reveals that it is possible. Nonetheless, when combining this information with the analysis of EAI s, we know that such a scenario is improbable. As far as expected efficiencies are concerned, they impose the following ranking on the set of DMUs: $D_1 > D_3 > D_2 > D_5 > D_4$, hence allowing discrimination between both efficient and inefficient units.

In the e-Appendix, we present a detailed step-by-step description of calculating the EAI s and other stochastic measures for the considered example. To make the description self-contained and its size reasonable, we use only ten samples as opposed 10,000 samples considered in the main paper.

2.3.2 Efficiency rank acceptability indices

Efficiency Rank Acceptability Index $ERAI(DMU_o, r)$ for $DMU_o \in \mathcal{D}$ and a specific rank $r \in \{1, 2, \dots, K\}$ is defined as the share of feasible scenarios for which DMU_o is placed at the r -th position in the ranking imposed by the efficiency scores of all DMUs in \mathcal{D} . Note that for each $DMU_o \in \mathcal{D}$, $\sum_{r=1}^K ERAI(DMU_o, r) = 1$. These stochastic indices can be used to approximate an expected efficiency rank (denoted by ER_o) for DMU_o in the following way: $ER_o = \sum_{r=1}^K r \cdot ERAI(DMU_o, r)$ (Ang et al. 2021). Similar to the expected efficiencies, the expected efficiency ranks can be used to order the units from the best to the worst.

Table 3 Stochastic results derived with the Monte Carlo simulation for the illustrative example

DMU_o	EAI_s					EE_o					$ERAIs$					ER_o	$PEOIs$					
	[0.0, 0.2]	(0.2, 0.4]	(0.4, 0.6]	(0.6, 0.8]	(0.8, 1.0]		1	2	3	4	5	1	2	3	4		5	D_1	D_2	D_3	D_4	D_5
D_1	0.000	0.002	0.008	0.032	0.958	0.980	0.83	0.17	0.00	0.00	0.00	0.00	0.83	0.17	0.00	0.00	1.17	1.00	1.00	0.83	1.00	1.00
D_2	0.182	0.716	0.102	0.000	0.000	0.289	0.00	0.00	0.66	0.34	0.00	0.00	0.66	0.00	0.00	0.00	3.34	0.00	1.00	0.00	1.00	0.66
D_3	0.000	0.081	0.101	0.440	0.378	0.749	0.17	0.83	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.83	0.17	1.00	1.00	1.00	1.00
D_4	0.790	0.210	0.000	0.000	0.000	0.134	0.00	0.00	0.00	0.16	0.84	0.00	0.00	0.00	0.00	0.00	4.84	0.00	0.00	0.00	1.00	0.16
D_5	0.220	0.694	0.086	0.000	0.000	0.272	0.00	0.00	0.34	0.50	0.16	0.00	0.34	0.00	0.00	0.00	3.82	0.00	0.34	0.00	0.84	1.00

Illustrative example The rank acceptabilities for the illustrative examples are presented in Table 3. Based on the derived samples' analysis, only D_1 and D_3 can be ranked in the first two positions. However, the probability of D_1 being ranked at the top is higher than for D_3 ($ERAI(D_1, 1) = 0.83 > ERAI(D_3, 1) = 0.17$). Even though the minimal rank for D_3 indicated that it could be ranked fourth in the most pessimistic case, the analysis of *ERAI*s suggests that the scenarios for which it drops out of the top two are very unlikely. The distribution of ranks for D_5 confirms that it is ranked fourth for half of the scenarios. Further, the probabilities of attaining the third and fifth positions by D_5 are equal to, respectively, 34% and 16%. The expected efficiency ranks impose the following order on the set of DMUs: $D_1 > D_3 > D_2 > D_5 > D_4$. Even though it exploits the ordinal results (i.e., ranks) rather than cardinal ones (i.e., efficiencies), this ranking is the same as when considering the expected efficiencies.

2.3.3 Pairwise efficiency outranking indices

The Pairwise Efficiency Outranking Index $PEOI(DMU_o, DMU_l)$ is defined as the share of feasible scenarios for which DMU_o is at least as efficient as DMU_l . Note that for $(DMU_o, DMU_l) \in \mathcal{D} \times \mathcal{D}$, $0 \leq PEOI(DMU_o, DMU_l) \leq 1$ and $0 \leq PEOI(DMU_o, DMU_l) + PEOI(DMU_l, DMU_o) \leq 2$.

Illustrative example The *PEOIs* derived for the illustrative example are presented in Table 3. Note that when for the pairs for which the necessary relation holds (e.g., (D_1, D_2) and (D_3, D_5)), *PEOI* is equal to one, whereas for the pairs for which the possible relation is false (e.g., (D_2, D_1) and (D_5, D_1)), *PEOI* is zero. The analysis of *PEOIs* is the most informative for pairs that are not related by the necessary relation. For example, the share of scenarios for which D_1 attains higher efficiency than D_3 is five times greater than the share for which the inverse relation holds. In the same spirit, D_2 is more efficient than D_5 for twice as many scenarios as D_5 being more favorable than D_2 . Having compared D_3 with D_2 or D_4 using the exact robust analysis methods, we know that these pairs are not related by \succ_E^N . However, *PEOIs* indicate that the scenarios for which D_2 and D_4 are strictly better than D_3 are extremely limited ($PEOI(D_2, D_3) = 0$ and $PEOI(D_4, D_3) = 0$).

To demonstrate the impact that joint consideration of variable weights and imprecise inputs and outputs has on the obtained robust results, in the e-Appendix, we reconsider the illustrative example. Specifically, we analyze five scenarios while replacing performances on a single or two imprecise factors with the respective precise data. For each scenario, we discuss the six types of results. In this way, we demonstrate that imprecision of inputs and outputs contributes to the uncertainty of efficiency outcomes in the same way as the multiplicity of weights associated with these factors.

3 Implementation on the *diviz* platform

Diviz is an open-source platform that allows designing and executing algorithmic workflows implementing operational research methods (Meyer and Bigaret 2012). The software consists of two major components: i) a Java client, which allows users to design workflows using existing computational and graphical modules, and ii) servers, where the computations are performed and the results are generated. The greatest number of contributions on *diviz* concern Multiple Criteria Decision Analysis (MCDA) (see Cinelli et al. 2022; Greco et al. 2016). All *diviz* modules take input data and produce outputs in XMCD, a dedicated XML-based format.

3.1 Implemented modules

All methods for robustness analysis in Imprecise DEA have been implemented in R and made available on the *diviz* platform as independent modules (web services). Their source code is available at https://github.com/alabijak/diviz_DEA/tree/master/ImpreciseDEACCR. They can be used individually or combined into complex workflows. Each module accepts five input files:

- *units* containing information about the considered DMUs;
- *inputs/outputs* listing information on the inputs and outputs and their scales (quantitative or qualitative (ordinal));
- *performance* providing information on the DMUs' precise performances or, if the problem involves interval inputs and outputs, the minimal performances of DMUs;
- *max performance* is an optional file used in case the interval inputs/outputs are considered; it defines the DMUs' maximal performances;
- *weights constraints* is an optional file containing linear constraints on the weights of inputs and outputs, defining the space of feasible weight vectors.

The modules admit the specification of some additional parameters. The most important ones are *samplesNo* indicating the number of samples derived with the Monte Carlo simulation realized with the HAR algorithm and *tolerance* (in %) used to convert the precise performances into interval ones. For example, a precise value x is transformed into the interval $[(1 - tolerance) \cdot x; (1 + tolerance) \cdot x]$.

The following modules for robustness analysis in IDEA have been implemented on *diviz*:

- *ImpreciseDEA-CCR_efficiencies* computes the minimal and maximal efficiencies (E^* and E_*) for each DMU using linear programming techniques;
- *ImpreciseDEA-CCR_extremeRanks* computes the best and the worst efficiency ranks (R^* and R_*) for each DMU using MILP;
- *ImpreciseDEA-CCR_preferenceRelations* verifies the truth of the necessary and possible efficiency preference relations for all pairs of DMUs using linear programming;

- *ImpreciseDEA-CCR-SMAA_efficiencies* computes the efficiency distribution, the extreme efficiency scores observed in the analyzed sample of feasible scenarios, and an expected efficiency for each DMU, using the HAR algorithm; it additionally requires specification of the number of buckets as a method parameter;
- *ImpreciseDEA-CCR-SMAA_preferenceRelations* computes *PEOIs* for all pairs of DMUs using HAR;
- *ImpreciseDEA-CCR-SMAA_ranks* computes *ERAI*s for all DMUs and ranks, extreme efficiency ranks observed in the analyzed sample of feasible scenarios, and an expected rank for each DMU using HAR.

The structures of two exemplary modules, *ImpreciseDEA-CCR_efficiencies* and *ImpreciseDEA-CCR-SMAA_efficiencies*, are presented in Figs. 1 and 2, respectively. They perform computations according to the methods presented in Sects. 2.2.1 and 2.3.1, respectively.

The implemented modules can be combined into an algorithmic workflow with other available computational or visualization modules. Such a workflow can be easily exported and shared with other users. Moreover, the infrastructure of *diviz* allows storing the history of past executions, which is very useful when comparing the results for different settings (e.g., with and without preference information specified by the user). The workflow designed to obtain the results for the case study discussed in Sect. 4 is graphically presented in Fig. 3.

4 Illustrative case study

To illustrate the practical usefulness of the proposed framework, we performed the robustness analysis for two studies concerning 27 industrial robots and 17 Chinese ports. The former is based on data derived from Saen (2006), and the detailed results are given in the e-Appendix. The latter builds on data from Jiang et al. (2021),

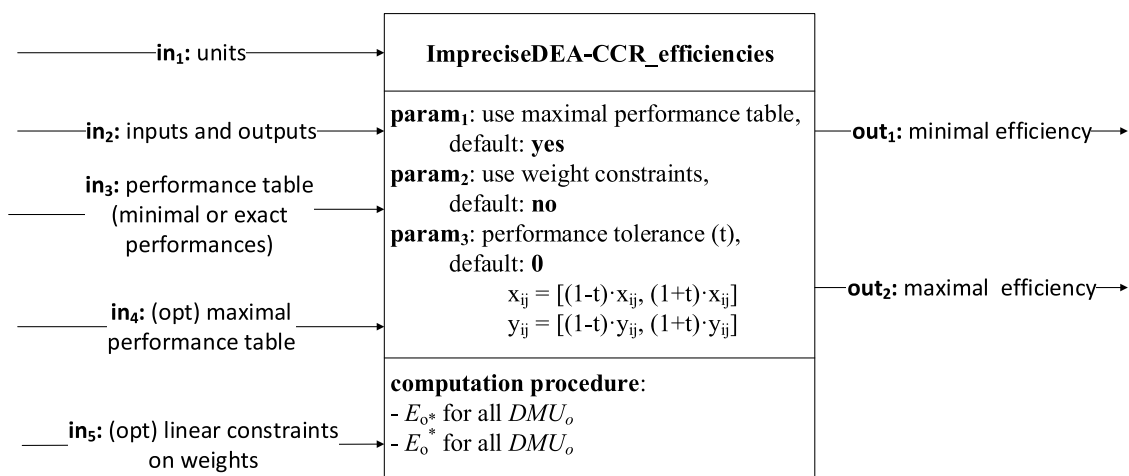


Fig. 1 The structure of the *diviz* module which computes the extreme efficiency scores for each DMU using MILP for the Imprecise DEA model

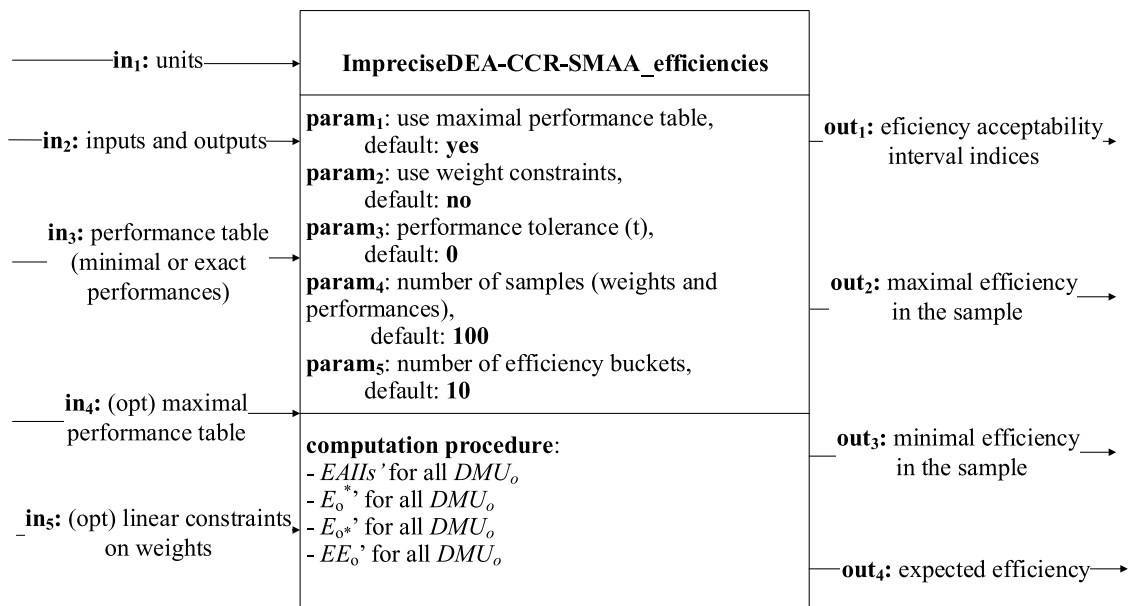


Fig. 2 The structure of the *diviz* module which computes the Efficiency Acceptability Interval Indices, observed extreme efficiency scores, and expected efficiency for each DMU using the Imprecise DEA model and the Monte Carlo simulation

and the outcomes are reported in this section. The workflows and input data in the XMCD format (ver. 2) for both studies are available at https://github.com/alabijak/diviz_DEA/tree/master/ImpreciseDEACCR.

The ports are described in terms of two precise inputs (labor population and energy consumption), two desirable outputs (cargo throughput—precise and employee satisfaction—ordinal), and one undesirable output (water pollutants—interval). Following (Jiang et al. 2021), the latter factor is treated as an input during the analysis. To obtain the same magnitude for all precise and interval factors, we used the mean normalization before running the methods (see Sarkis 2007; Tomažević et al. 2016; Widiarto and Emrouznejad 2015). The performances of ports on all inputs and outputs are presented in Table 4.

In what follows, we discuss the results of robustness analysis considering the three perspectives: efficiency scores, efficiency ranks, and preference relations. The values of stochastic acceptability indices are estimated based on the 10,000 sampled feasible scenarios. To illustrate that the methods can handle linear weight constraints, we assess water pollutants as less important factor than the other two inputs, i.e., $u_{wp} \leq u_{lp}$ and $u_{wp} \leq u_{ec}$, where u_{wp}, u_{lp}, u_{ec} are, respectively, weights assigned to water pollutants, labor population, and energy consumption. Moreover, we introduce two other constraints preventing the overwhelming role of any input, i.e., $w_{lp} \leq w_{ec} + w_{wp}$ and $w_{ec} \leq w_{lp} + w_{wp}$.

4.1 Efficiency scores

Figure 4 presents the extreme efficiency scores (E^* and E_*) for all DMUs. Regarding the maximal (optimistic) efficiencies, they indicate six efficient ports (Yingkou, Tianjin, Yantai, Ningbo-zhoushan, Fuzhou, and Shantou) with $E^* = 1$. Among the

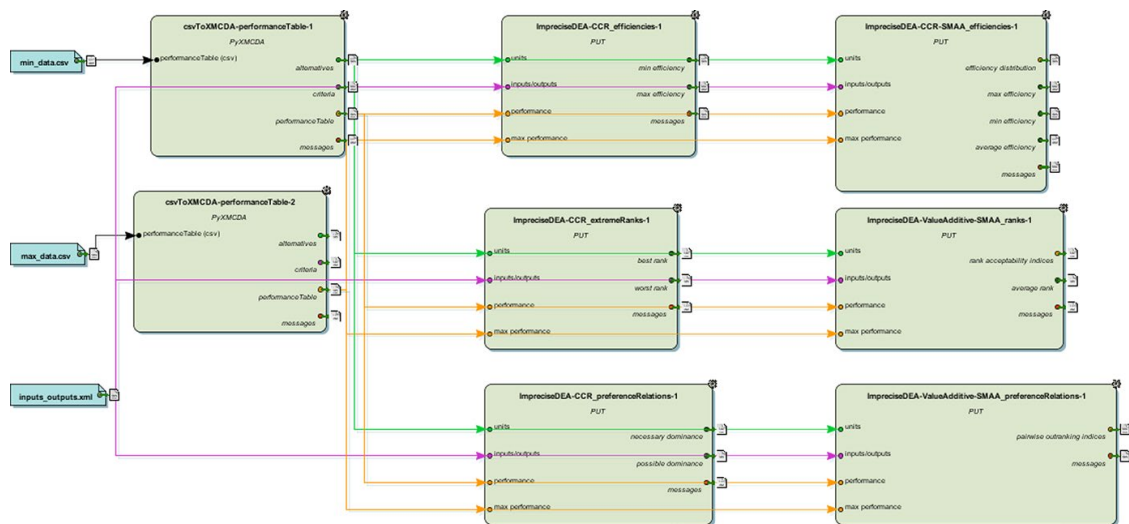


Fig. 3 The *diviz* workflow used to perform the efficiency analysis for a case study (see Sect. 4)

inefficient ports, the best efficiency is attained by Fangcheng (0.909) and Zhanjiang (0.887). On the other extreme, there are two ports with maximal efficiency scores lower than 0.6 (Shanghai (0.539) and Qinhuangdao (0.598)). When analyzing the minimal (pessimistic) efficiencies, the most advantageous port is Ningbo-zhoushan (0.158). The minimal efficiency scores of all other ports are significantly lower (close to zero).

In general, the efficiency intervals are relatively wide. For this reason, analyzing the distribution of efficiency scores is desirable. In Table 5, we present the Efficiency Acceptability Interval Indices, while assuming $B = 10$ sub-intervals from $[0, 0.1]$ to $(0.9, 1.0]$. When it comes to the efficient ports, the greatest *EAI* for the best interval is attained by Tianjin ($EAI(\text{Tianjin}, (0.9, 1.0]) = 57.3\%$) and Yantai ($EAI(\text{Yantai}, (0.9, 1.0]) = 68.3\%$). Only three other ports attained an efficiency greater than 0.9 for at least one sample, but the respective *EAI*s are significantly lower (16.8% for Shantou and less than 9% for others). Interestingly, Fuzhou—deemed efficient—has not achieved an efficiency score in the best interval for any weight sample. Obviously, such scores are feasible (as confirmed with the analysis of exact extreme scores), but *EAI*s indicate that they are improbable.

For some ports, the analysis of *EAI*s allows indicating the most probable ranges of efficiencies even if the efficiency intervals are relatively wide. For example, the efficiency score for Yantai is in the best three buckets for 98.6% of feasible scenarios, with the vast majority (68.3%) in the last bucket. In the same spirit, the efficiency score of Qinhuangdao is between 0.2 and 0.4 for 89.8% of feasible scenarios, and there is no sample for which its efficiency is greater than 0.5. However, there is also a group of ports with efficiency scores strongly dependent on the selected weight and performance vectors. For example, for Shantou, *EAI*s greater than 16% are attained for the two very different intervals, $(0.9, 1]$ and $(0.3, 0.4]$. Also, for this port and eight buckets representing efficiency scores between 0.2 and 1.0, *EAI*s are greater than zero. Similarly, Fuzhou has a positive share of feasible scenarios for nine sub-intervals.

Table 4 Input and output values for considered Chinese ports (for employee satisfaction, 1 and 17 indicate, respectively, the worst and the best ordinal performances)

Port	Inputs		Outputs		Undesirable output
	Labor population	Energy consumption	Cargo throughput	Employee satisfaction	Water pollutants
Dalian	0.961	0.997	1.255	15	[1.137, 1.421]
Yingkou	0.641	1.093	1.012	16	[0.796, 1.023]
Qinhuangdao	1.827	0.406	0.537	9	[0.455, 0.682]
Tianjin	1.022	0.533	1.583	15	[1.251, 1.706]
Yantai	0.208	0.798	0.763	15	[0.569, 0.853]
Qingdao	1.686	1.701	1.438	15	[1.137, 1.706]
Rizhao	0.763	0.830	1.006	9	[0.910, 1.137]
Shanghai	2.524	2.811	1.854	11	[1.706, 2.161]
Lianyungang	0.574	0.721	0.577	9	[0.455, 0.682]
Ningbo-zhoushan	1.895	2.533	2.651	17	[1.990, 2.843]
Fuzhou	0.434	0.465	0.417	6	[0.341, 0.455]
Xiamen	0.980	0.600	0.601	6	[0.398, 0.569]
Shantou	0.601	0.141	0.143	11	[0.284, 0.398]
Shenzhen	0.989	0.444	0.615	1	[0.512, 0.625]
Guangzhou	1.096	1.499	1.502	6	[1.706, 1.990]
Zhanjiang	0.353	0.971	0.736	6	[0.569, 0.682]
Fangcheng	0.447	0.456	0.307	6	[0.341, 0.512]

The distribution of efficiency scores can be translated into a single, easily understandable measure, i.e., expected efficiency (see Fig. 4). These scores impose a complete order on the set of ports. Yantai is the best, with an expected efficiency of 0.930. This means it is either efficient or very close to being efficient for most feasible scenarios. The other two ports in the top three are Tianjin (0.877) and Yingkou (0.768). Dalian attains the next highest expected efficiency. Even though it is inefficient, it is ranked better in terms of *EE* than the remaining three efficient ports (Ningbo-zhoushan, Fuzhou, and Shantou). The three ports with the least expected efficiencies are Shenzhen (0.376), Shanghai (0.334), and Qinhuangdao (0.307).

4.2 Efficiency ranks

The extreme efficiency ranks (R^* and R_*) for all ports are presented in Fig. 5. Only the six ports deemed as efficient have the best rank equal to one. Furthermore, the inefficient units with relatively high maximal efficiency scores attain the best rank equal to two (see Dalian, Rizhao, Zhanjiang, and Fangcheng). Only one additional inefficient port (Lianyungang) is ranked at the podium in the best case. Four ports are always ranked outside the top five (see Shanghai, Xiamen,

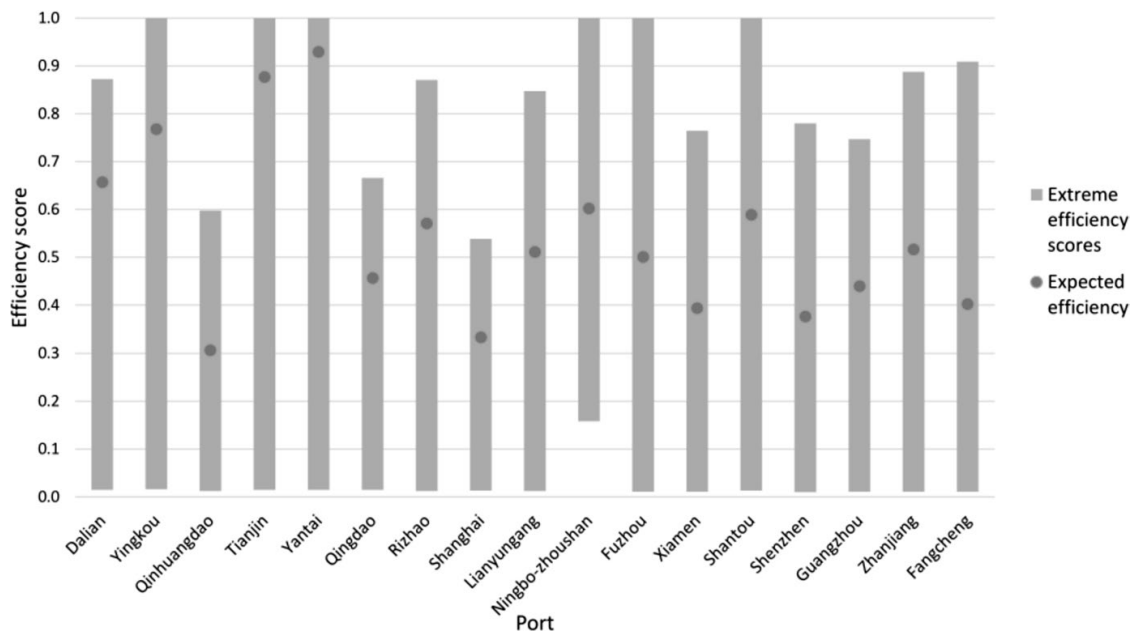


Fig. 4 Graphical representation of extreme efficiency scores and expected efficiencies for seventeen ports

Shenzhen, and Guangzhou). The least advantageous among them is Shanghai because even in the optimistic scenario, as many as 7 other ports attain higher efficiency ($R^* = 8$).

The most preferred ports in terms of the worst efficiency ranks are Yantai ($R_* = 7$) and Yingkou ($R_* = 11$). When it comes to other efficient ports, their ranks drop to 14-th (for Tianjin), 16-th (for Ningbo-zhoushan and Fuzhou), or 17-th (for Shantou) positions in the most pessimistic case. Hence the ranks of Shantou range between the most extreme possible ones. In general, for 13 out of 17 ports, the pessimistic rank is lower than 15-th. Among them, six inefficient ports (Qinhuangdao, Shanghai, Xiamen, Shenzhen, Guangzhou, and Fangcheng) are ranked at the bottom in the least advantageous scenario.

Since the rank intervals for all ports are broad, we have estimated Efficiency Rank Acceptability Indices (see Table 6). They reveal the distribution of ranks attained by each DMU across the feasible weight and performance vectors. Four of six efficient ports attained the top position for at least one feasible scenario. For Yantai, this occurs for 58.9% of samples, Tianjin and Shantou are ranked at the top for similar shares of scenarios (19.6% and 15.3%, respectively), whereas for Yingkou, the *ERAI* for the first position is significantly lesser (6.2%). Regarding the remaining efficient ports, Ningbo-zhoushan attained at most second rank, and Fuzhou was at most third for relatively negligible shares of feasible scenarios.

For many ports, it is possible to indicate a single rank or a relatively narrow range of ranks attained for the vast majority of feasible scenarios. For example, Yingkou is ranked third for more than 60% scenarios, Yantai is ranked in the top three for all sampled scenarios, and Dalian is placed between fourth and sixth for more than 85% scenarios. Some other ports attain a more extensive range of ranks for a significant

Table 5 Efficiency distribution for considered ports obtained with the Monte Carlo simulation

Port	[0.0, 0.1]	(0.1, 0.2]	(0.2, 0.3]	(0.3, 0.4]	(0.4, 0.5]	(0.5, 0.6]	(0.6, 0.7]	(0.7, 0.8]	(0.8, 0.9]	(0.9, 1.0]
Dalian	0.000	0.000	0.002	0.007	0.052	0.154	0.443	0.335	0.007	0.000
Yingkou	0.000	0.000	0.000	0.000	0.001	0.021	0.221	0.447	0.221	0.089
Qinhuangdao	0.002	0.047	0.390	0.508	0.053	0.000	0.000	0.000	0.000	0.000
Tianjin	0.000	0.000	0.002	0.000	0.015	0.056	0.090	0.113	0.151	0.573
Yantai	0.000	0.000	0.000	0.000	0.001	0.001	0.012	0.093	0.210	0.683
Qingdao	0.000	0.002	0.039	0.168	0.475	0.309	0.007	0.000	0.000	0.000
Rizhao	0.000	0.012	0.040	0.084	0.140	0.169	0.426	0.127	0.002	0.000
Shanghai	0.004	0.112	0.203	0.411	0.269	0.001	0.000	0.000	0.000	0.000
Lianyungang	0.000	0.002	0.026	0.089	0.262	0.483	0.135	0.003	0.000	0.000
Ningbo-zhoushan	0.000	0.000	0.010	0.064	0.115	0.196	0.444	0.163	0.007	0.001
Fuzhou	0.003	0.019	0.040	0.114	0.221	0.446	0.144	0.012	0.001	0.000
Xiamen	0.011	0.050	0.135	0.240	0.436	0.126	0.002	0.000	0.000	0.000
Shantou	0.000	0.000	0.139	0.165	0.132	0.118	0.111	0.085	0.082	0.168
Shenzhen	0.043	0.091	0.144	0.185	0.359	0.175	0.003	0.000	0.000	0.000
Guangzhou	0.022	0.086	0.107	0.141	0.166	0.351	0.127	0.000	0.000	0.000
Zhanjiang	0.004	0.033	0.068	0.112	0.155	0.303	0.255	0.070	0.000	0.000
Fangcheng	0.004	0.026	0.081	0.350	0.429	0.092	0.016	0.002	0.000	0.000

share of feasible scenarios. For example, Shantou has non-zero *ERAI* for all ranks, with the acceptability indices ranging from 1.2% to 18.6%. Furthermore, Rizhao and Lianyungang have *ERAI*s greater than 10% for five consecutive ranks. Also, the analysis of *ERAI*s leads to identifying some ranks which are feasible according to the exact robustness analysis while being less probable, as confirmed by the stochastic analysis conducted with the Monte Carlo simulation. For example, Qinhuangdao could be ranked fifth in the best case, but the probabilities of ranks above 8 are already close to zero. In the same spirit, Tianjin could be ranked 14-th in the worst case, but the shares of scenarios ranking it worse than 9 are negligible. In general, the analysis of *ERAI*s points out the ports for which the attained ranks are rather stable, or the ranks' variability is great. This means that their position strongly depends on the particular scenario (weights and performances).

*ERAI*s can be transformed into expected ranks that allow ordering the ports from the best to the worst (see Fig. 5). In particular, for Yantai, *ER* is equal to 1.438, which confirms its superiority over the remaining ports. The following two positions are attained by Tianjin (2.827) and Yingkou (2.918). Even though Dalian is inefficient, its expected rank is better than those attained by the remaining three efficient ports. Qinhuangdao, Shanghai, and Shenzhen attain the worst expected ranks. They confirm that these three ports are placed at the bottom for most feasible scenarios.

4.3 Preference relations

The third analyzed perspective concerns the stability of efficiency-based pairwise preference relations. In Table 7, we present the matrix summarizing the truth of the necessary (N) and possible (P) preference relations. First, let us note that the necessary relation is reflexive, which is confirmed with “N” on the main diagonal of Table 7. Furthermore, the necessary relation holds for 20 pairs involving different

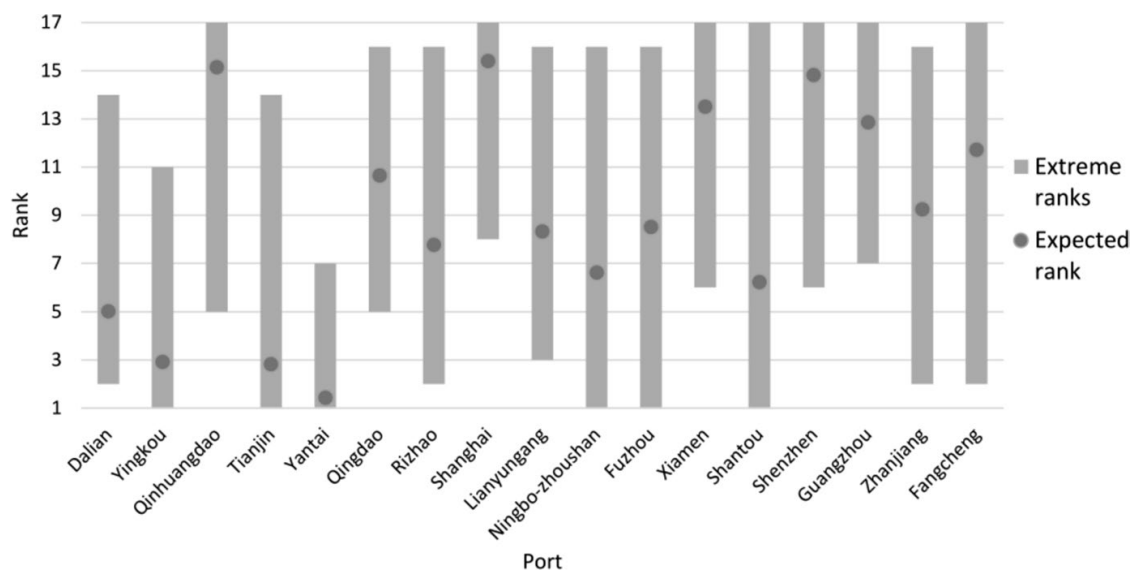


Fig. 5 Graphical representation of extreme efficiency ranks and estimated expected ranks (note that the closer to the bottom of the figure, the better the ranks)

Table 6 Efficiency Rank Acceptability Indices for analyzed ports obtained with the Monte Carlo simulation

Port	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Dalian	0.000	0.000	0.046	0.290	0.431	0.133	0.045	0.036	0.017	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000
Yingkou	0.062	0.169	0.613	0.119	0.023	0.010	0.004	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Qinhuangdao	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.005	0.008	0.051	0.022	0.039	0.071	0.123	0.110	0.189	0.382
Tianjin	0.196	0.260	0.174	0.303	0.040	0.017	0.006	0.003	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Yantai	0.589	0.384	0.027	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Qingdao	0.000	0.000	0.000	0.000	0.000	0.014	0.092	0.062	0.114	0.095	0.260	0.198	0.102	0.055	0.008	0.000	0.000
Rizhao	0.000	0.000	0.005	0.027	0.086	0.182	0.211	0.142	0.125	0.103	0.076	0.037	0.006	0.000	0.000	0.000	0.000
Shanghai	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.004	0.002	0.005	0.013	0.018	0.039	0.141	0.214	0.336	0.228
Lianyungang	0.000	0.000	0.000	0.015	0.039	0.113	0.150	0.215	0.176	0.203	0.067	0.021	0.001	0.000	0.000	0.000	0.000
Ningbo-zhoushan	0.000	0.001	0.025	0.078	0.224	0.300	0.111	0.078	0.064	0.039	0.046	0.021	0.005	0.008	0.000	0.000	0.000
Fuzhou	0.000	0.000	0.004	0.024	0.061	0.090	0.107	0.163	0.212	0.169	0.132	0.034	0.004	0.000	0.000	0.000	0.000
Xiamen	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.032	0.103	0.373	0.329	0.133	0.022	0.006
Shantou	0.153	0.186	0.100	0.122	0.057	0.051	0.023	0.025	0.024	0.020	0.019	0.039	0.012	0.016	0.022	0.070	0.061
Shenzhen	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.003	0.007	0.024	0.128	0.105	0.099	0.218	0.168	0.248
Guangzhou	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.128	0.059	0.058	0.061	0.122	0.083	0.108	0.145	0.170	0.064
Zhanjiang	0.000	0.000	0.004	0.015	0.023	0.056	0.201	0.085	0.126	0.187	0.122	0.094	0.063	0.023	0.001	0.000	0.000
Fangcheng	0.000	0.000	0.002	0.007	0.016	0.034	0.048	0.054	0.069	0.060	0.125	0.146	0.136	0.098	0.149	0.045	0.011

ports. This means that for all feasible scenarios, one port attains efficiency at least as good as the other port, confirming the robustness of its advantage. For example, Tianjin is necessarily preferred to Qingdao, Guangzhou, and Shanghai.

The necessary relation can be presented graphically using the Hasse diagram (see Fig. 6). Such a diagram does not represent the truth of relations that can be derived from the transitivity. For example, since Dalian is necessarily preferred to Qingdao, and Qingdao is necessarily preferred to Shanghai, Dalian is also preferred to Shanghai. Note that no other port is necessarily preferred over six efficient ports. However, there are also four inefficient (Dalian, Rizhao, Shenzhen, and Fangcheng) for which there does not exist any other port confirming its necessary advantage over them.

The port which proves its robust superiority over the greatest number of other ports is Yantai. It is necessarily preferred over seven other ports. An interesting situation can be observed for Fuzhou and Shantou. Although they are efficient, they are not necessarily preferred over any other port. The latter (i.e., no outgoing arc) also holds for seven other ports. Among them, Shanghai and Guangzhou are necessarily worse than the most significant number of other ports (6 and 5, respectively). The necessary relation graph suggests the improvement paths that inefficient ports can follow to improve their efficiency gradually. For example, for Shanghai, we can construct the following example paths: Shanghai–Qingdao–Dalian or Shanghai–Qingdao–Yingkou. Alternatively, it can directly learn from Ningbo-zhoushan.

Regarding the possible preference relation (see Table 7), let us emphasize that the necessary relation implies the truth of the possible one. However, the latter one holds also for pairs that are not related by the necessary preference. For example, all efficient ports are incomparable in terms of the necessary relation. This means there is at least one feasible scenario for which one port is preferred to the other and at least one feasible scenario for which this relation is inverse (e.g., Yingkou and Tianjin). Such a situation also occurs for pairs of inefficient ports (e.g., Dalian and Rizhao or Qingdao and Guangzhou). When the possible relation for a given pair of ports is false (e.g., (Qingdao, Dalian) or (Shanghai, Qingdao)), then one port is less efficient than the other for all feasible scenarios.

When the necessary relation is true or the possible relation is false, a pair of ports is compared in the same way for all feasible scenarios. However, when the ports are incomparable in terms of the necessary relation, it is interesting to analyze the shares of feasible scenarios that rank one of the ports at least as good as the other. Pairwise Efficiency Outranking Indices capture such shares (see Table 8).

For some pairs, these indices confirm the superiority of one port over the other. For example, $PEOI(\text{Yingkou}, \text{Zhanjiang}) = 0.995$ and $PEOI(\text{Rizhao}, \text{Qinhuangdao}) = 0.999$ indicate that one port is at least as efficient as the other for over 99% scenarios, and the inverse relation is extremely unlikely with $PEOIs$ close to zero. For some other pairs, the $PEOIs$ indicate that the shares of scenarios confirming the advantage of either port over another are very similar (e.g., $PEOI(\text{Shanghai}, \text{Qinhuangdao}) = 0.506$ and $PEOI(\text{Qinhuangdao}, \text{Shanghai}) = 0.494$ or $PEOI(\text{Shenzhen}, \text{Qinhuangdao}) = 0.540$ and $PEOI(\text{Qinhuangdao}, \text{Shenzhen}) = 0.460$). For these pairs, the indication of a more preferred port strongly depends on the selected weights and performance vectors.

Table 7 The necessary and possible efficiency preference relations for all pairs of ports (N—necessary relation; P—possible relation; empty cell—the relation is not even possible)

Port	Dalian	Yingkou	Qinhuangdao	Tianjin	Tianjindao	Qingdao	Rizhao	Shanghai	Lianyungang	Ningbo-zhoushan	Fuzhou	Xiamen	Shantou	Shenzhen	Guangzhou	Zhanjiang	Fangcheng
Dalian	N	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
Yingkou	P	N	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
Qinhuangdao	P	P	N	P	P	P	P	P	P	P	P	P	P	P	P	P	P
Tianjin	P	P	P	N	P	P	P	P	P	P	P	P	P	P	P	P	P
Yantai	P	P	P	P	N	P	P	P	P	P	P	P	P	P	P	P	P
Qingdao	P	P	P	P	P	N	P	P	P	P	P	P	P	P	P	P	P
Rizhao	P	P	P	P	P	P	N	P	P	P	P	P	P	P	P	P	P
Shanghai	P	P	P	P	P	P	P	N	P	P	P	P	P	P	P	P	P
Lianyungang	P	P	P	P	P	P	P	P	N	P	P	P	P	P	P	P	P
Ningbo-zhoushan	P	P	P	P	P	P	P	P	P	N	P	P	P	P	P	P	P
Fuzhou	P	P	P	P	P	P	P	P	P	P	N	P	P	P	P	P	P
Xiamen	P	P	P	P	P	P	P	P	P	P	P	N	P	P	P	P	P
Shantou	P	P	P	P	P	P	P	P	P	P	P	P	N	P	P	P	P
Shenzhen	P	P	P	P	P	P	P	P	P	P	P	P	P	N	P	P	P
Guangzhou	P	P	P	P	P	P	P	P	P	P	P	P	P	P	N	P	P
Zhanjiang	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	N	P
Fangcheng	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	N

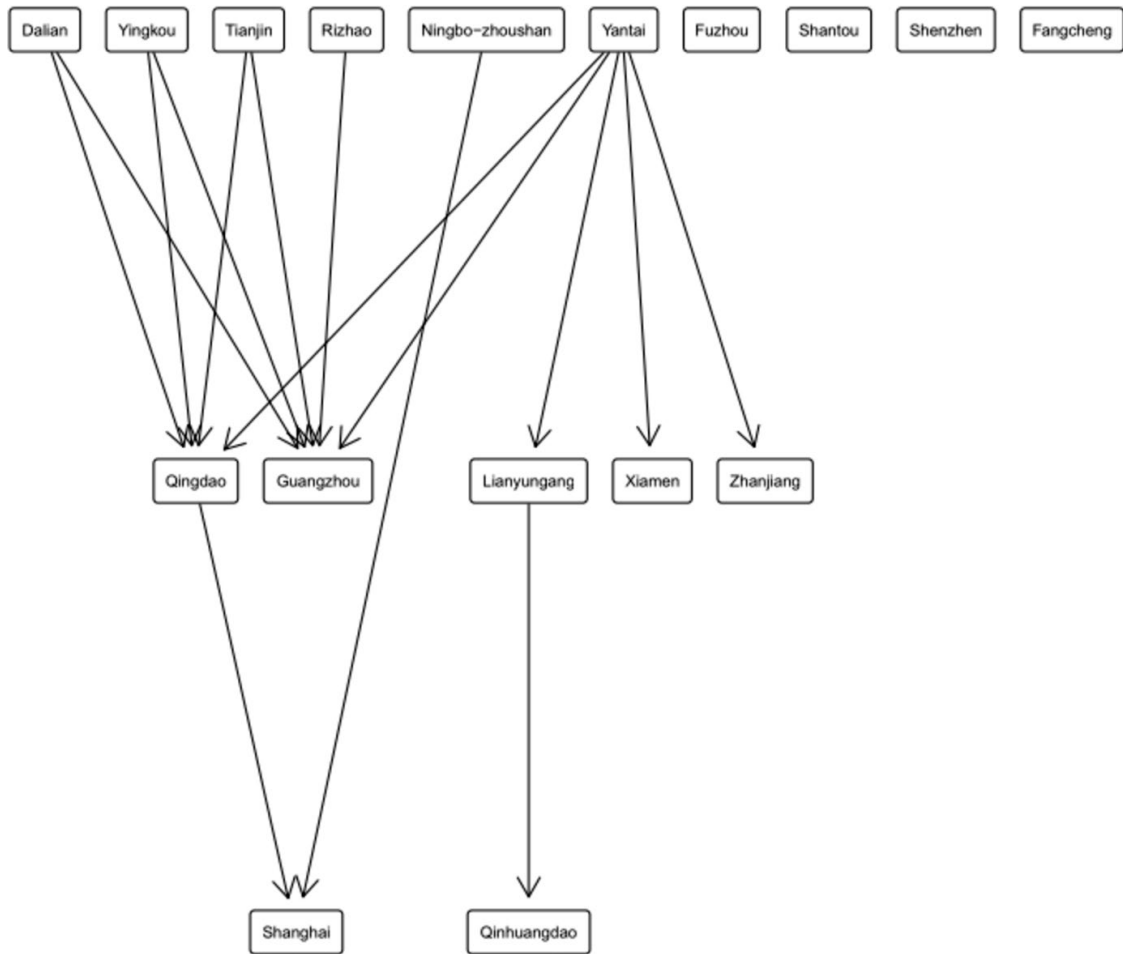


Fig. 6 The Hasse diagram representing the necessary efficiency preference relation

5 Conclusions and future work

We have introduced a rich framework for robustness in the context of Imprecise Data Envelopment Analysis. The proposed methods are applicable in the context of three major types of uncertainty that occur in real-world decision problems (Pelissari et al. 2021). First, we consider ambiguity in the input and output performances that could be interpreted differently due to their ordinal or interval character. Second, we account for stochasticity by considering discrete and continuous probability distributions. Third, we deal with partial information on the input and output weights by exploiting the space of feasible weights delimited with a set of linear constraints. In this way, the proposed approaches for robustness analysis can be applied to real-world problems for which it is difficult to express the knowledge or collect precise data, the variables are unquantifiable, some errors in measurements occur, or the users are not able or willing to express their complete preferences (see Dehnokhalaji et al. 2022; Pelissari et al. 2021).

When considering the stability of results that can be attained for the possible performances and weights, we focus on three types of efficiency-based outcomes: scores, preference relations, and ranks. On the one hand, the mathematical

programming models compute the extreme efficiency scores and ranks and verify the truth of the necessary and possible preference relations. These outcomes reveal each DMU's performance for the most and the least advantageous scenarios and collate the efficiencies of all pairs of DMUs for all or at least one feasible scenario. Thus, they offer an exact perspective on the DMUs' performance. On the other hand, we incorporate the stochastic analysis driven by the Monte Carlo simulations to derive the probability distribution of different outcomes and expected results. These stochastic outcomes complement the exact results derived using mathematical programming. They also provide means for analyzing trends or some prevailing scenarios and imposing the ranking on the set of DMUs in the line of expected scores or ranks. Above all, such various outcomes have not been offered by any previous IDEA method.

The practical usefulness of the proposed framework has been illustrated in real-world case studies concerning the evaluation of Chinese ports (Jiang et al. 2021) and industrial robots (Saen 2006). The data sets involved precise, interval, and ordinal factors. The results were computed with modules implemented in R and available on the *diviz* platform (Meyer and Bigaret 2012). They incorporate MILP solver and advanced sampling techniques.

The main limitations of the proposed framework are three-fold. First, when the number of units runs over a few hundred, the linear programs are too big and too many, posing a significant problem for contemporary solvers. This is particularly true for results such as extreme ranks that are established using binary variables. Moreover, for such big data problems, the robust results, such as the ranking induced by the necessary preference relation, cannot be presented to the user because of their high complexity. Then, it is more beneficial to refer to complete orders of units based on expected efficiencies or ranks. Still, let us emphasize that large scale-applications are less common in DEA. Second, the results of the stochastic analysis depend on the assumed distribution of weights and performances within the intervals as well as the hypotheses made when representing the ordinal performances. Clearly, the developed framework is applicable with other types of distributions than uniform and arbitrary performance ranges from which the ordinal performances could be sampled. These choices may affect the values of stochastic acceptabilities. Also, whichever assumptions are made, even if the indices can be estimated within the acceptable error bound, they are not accurate. Third, we accounted for the standard imprecision types considered in IDEA, including interval and ordinal performances. As proved by a comprehensive review by Pelissari et al. (2021), uncertain performances can also be modeled differently. The most popular approaches for this purpose include fuzzy numbers, non-uniform probabilistic distributions, evidential reasoning, and grey numbers. Their combinations with DEA have gained in popularity in recent years. Hence adopting the framework for robustness analysis to their context remains an appealing direction for future research.

We develop the methods introduced in this paper in the following directions. First, we adapt them to hierarchical structures of inputs and outputs. In this way, the robustness of efficiency outcomes given imprecise performances can be investigated at the comprehensive and local levels (Shen et al. 2013). The latter corresponds to a more elementary perspective or particular sub-area of DMUs' functioning. Indeed,

Table 8 Pairwise Efficiency Outranking Indices for all pairs of ports obtained with the Monte Carlo simulation

Port	Dalian	Yingkou	Qinhuangdao	Tianjin	Yantai	Qingdao	Rizhao	Shanghai	Lianyungang	Ningbo	Fuzhou	Xiamen	Shantou	Shenzhen	Guangzhou	Zhanjiang	Fangcheng
Dalian	1.000	0.069	1.000	0.002	0.000	1.000	0.979	1.000	0.910	0.834	0.893	1.000	0.410	1.000	1.000	0.967	0.937
Yingkou	0.931	1.000	1.000	0.499	0.069	1.000	0.968	1.000	1.000	0.970	0.990	1.000	0.598	1.000	1.000	0.995	0.995
Qinhuangdao	0.000	0.000	1.000	0.000	0.000	0.052	0.001	0.494	0.000	0.020	0.058	0.218	0.062	0.460	0.385	0.118	0.092
Tianjin	0.998	0.501	1.000	1.000	0.207	1.000	1.000	1.000	0.981	0.973	0.968	1.000	0.552	1.000	1.000	0.999	0.973
Yantai	1.000	0.931	1.000	0.793	1.000	1.000	1.000	1.000	1.000	0.996	1.000	1.000	0.809	1.000	1.000	1.000	1.000
Qingdao	0.000	0.000	0.948	0.000	0.000	1.000	0.194	1.000	0.204	0.018	0.300	0.881	0.236	0.944	0.721	0.349	0.643
Rizhao	0.021	0.032	0.999	0.000	0.000	0.806	1.000	0.995	0.511	0.307	0.645	0.990	0.318	1.000	1.000	0.781	0.759
Shanghai	0.000	0.000	0.506	0.000	0.000	0.000	0.005	1.000	0.006	0.000	0.029	0.157	0.146	0.412	0.210	0.048	0.218
Lianyungang	0.090	0.000	1.000	0.019	0.000	0.796	0.489	0.994	1.000	0.280	0.484	0.999	0.272	0.993	0.815	0.602	0.840
Ningbo	0.166	0.030	0.980	0.027	0.004	0.982	0.693	1.000	0.720	1.000	0.763	0.966	0.376	0.997	1.000	0.856	0.839
Fuzhou	0.107	0.010	0.942	0.032	0.000	0.700	0.355	0.971	0.516	0.237	1.000	1.000	0.274	0.998	0.808	0.558	0.981
Xiamen	0.000	0.000	0.782	0.000	0.000	0.119	0.010	0.843	0.001	0.034	0.000	1.000	0.184	0.717	0.487	0.004	0.284
Shantou	0.590	0.402	0.938	0.448	0.191	0.764	0.682	0.854	0.728	0.624	0.726	0.816	1.000	0.812	0.758	0.711	0.830
Shenzhen	0.000	0.000	0.540	0.000	0.000	0.056	0.000	0.588	0.007	0.003	0.002	0.283	0.188	1.000	0.174	0.000	0.266
Guangzhou	0.000	0.000	0.615	0.000	0.000	0.279	0.000	0.790	0.185	0.000	0.192	0.513	0.242	0.826	1.000	0.004	0.418
Zhanjiang	0.033	0.005	0.882	0.001	0.000	0.651	0.219	0.952	0.398	0.144	0.442	0.996	0.289	1.000	0.996	1.000	0.618
Fangcheng	0.063	0.005	0.908	0.027	0.000	0.357	0.241	0.782	0.160	0.161	0.019	0.716	0.170	0.734	0.582	0.382	1.000

the hierarchical structure is helpful for decomposing complex decision problems into smaller, manageable sub-problems. This is particularly useful in scenarios with high numbers of inputs and outputs, which often happens in medicine, energy, banking, and finances. Moreover, we design the Multiple Objective Optimization algorithms to determine different scenarios of improvements (i.e., reductions of consumed resources or increases of produced results) required for attaining or maintaining a particular target. These targets refer to many robust results, e.g., being necessarily ranked in the top three or attaining an efficiency score of at least 0.7 for all feasible scenarios (Ciomek et al. 2018). Moreover, the proposed framework offers flexibility to the Decision Maker, who can indicate which factors should be modified and to which extent. The obtained solutions reflect the trade-offs between modifications needed on various factors. Their analysis may lead to selecting the most preferred solution to be implemented in practice. Finally, we develop the methods for robustness analysis in the context of value-based DEA (see Gouveia et al. 2008; Labijak-Kowalska et al. 2023). This model is based on concepts from Multiple Criteria Decision Analysis, allowing the incorporation of managerial preferences on different levels. In this regard, the uncertainty is related to performances, weights, and the shape of value functions for inputs and outputs. However, the output types produced by these methods are similar to those discussed in this paper.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s12351-023-00755-z>.

Author contributions AL-K: Conceptualization, Methodology, Software, Investigation, Data curation, Writing- Original draft, Writing-review and editing, Visualization. MK: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing- Original draft, Writing-review and editing, Supervision, Project administration, Funding acquisition.

Funding The research of Anna Labijak-Kowalska was supported by the Polish Ministry of Education and Science, grant no. 0311/SBAD/0726. Miłosz Kadziński acknowledges support from the Polish National Science Center under the SONATA BIS project (Grant No. DEC-2019/34/E/HS4/00045).

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Adler N, Friedman L, Sinuany-Stern Z (2002) Review of ranking methods in the Data Envelopment Analysis context. *Eur J Oper Res* 140(2):249–265
- Ang S, Zhu Y, Yang F (2021) Efficiency evaluation and ranking of supply chains based on stochastic multicriteria acceptability analysis and Data Envelopment Analysis. *Int Trans Oper Res* 28:3190–3219
- Aparicio J, Ruiz JL, Sirvent I (2007) Closest targets and minimum distance to the Pareto-efficient frontier in DEA. *J Prod Anal* 28(3):209–218
- Aparicio J, Cordero JM, Ortiz L (2019) Measuring efficiency in education: the influence of imprecision and variability in data on DEA estimates. *Socioecon Plan Sci* 68:100698
- Azadi M, Saen RF (2013) A combination of QFD and imprecise DEA with enhanced Russell graph measure: a case study in healthcare. *Socioecon Plan Sci* 47(4):281–291
- Azizi H, Kordrostami S, Amirteimoori A (2015) Slacks-based measures of efficiency in imprecise Data Envelopment Analysis: an approach based on Data Envelopment Analysis with double frontiers. *Comput Ind Eng* 79:42–51
- Charnes A, Cooper WW, Rhodes E (1978) Measuring the efficiency of decision making units. *Eur J Oper Res* 2(6):429–444
- Charnes A, Cooper W, Lewin A, Seiford L (1994) *Data Envelopment Analysis: theory, methodology, and applications*. Springer, Netherlands
- Chen L, Wang Y-M (2020) DEA target setting approach within the cross efficiency framework. *Omega* 96:102072
- Cinelli M, Kadziński M, Miebs G, Gonzalez M, Słowiński R (2022) Recommending multiple criteria decision analysis methods with a new taxonomy-based decision support system. *Eur J Oper Res* 302(2):633–65
- Ciomek K, Kadziński M (2021) Polyrun: a Java library for sampling from the bounded convex polytopes. *SoftwareX* 13:100659
- Ciomek K, Ferretti V, Kadziński M (2018) Predictive analytics and disused railways requalification: insights from a Post Factum Analysis perspective. *Decis Support Syst* 105:34–51
- Cook WD, Seiford LM (2009) Data Envelopment Analysis (DEA)—thirty years on. *Eur J Oper Res* 192(1):1–17
- Cooper WW, Park KS, Yu G (1999) IDEA and AR-IDEA: models for dealing with imprecise data in DEA. *Manag Sci* 45(4):597–607
- Cooper WW, Park KS, Yu G (2001) An illustrative application of IDEA (Imprecise Data Envelopment Analysis) to a Korean mobile telecommunication company. *Oper Res* 49(6):807–820
- Cooper W, Seiford L, Zhu J (2014) *Handbook on Data Envelopment Analysis*. International series in operations research & management science. Springer, New York
- Corrente S, Greco S, Słowiński R (2017) Handling imprecise evaluations in multiple criteria decision aiding and robust ordinal regression by n-point intervals. *Fuzzy Optim Decis Mak* 16(2):127–157
- Dehnokhalaji A, Khezri S, Emrouznejad S (2022) A box-uncertainty in DEA: a robust performance measurement framework. *Expert Syst Appl* 187:115855
- Despotis DK, Smirlis YG (2002) Data Envelopment Analysis with imprecise data. *Eur J Oper Res* 140(1):24–36
- Ebrahimi B, Khalili M (2018) A new integrated AR-IDEA model to find the best DMU in the presence of both weight restrictions and imprecise data. *Comput Ind Eng* 125:357–363
- Ebrahimi B, Toloo M (2020) Efficiency bounds and efficiency classifications in imprecise DEA: an extension. *J Oper Res Soc* 71(3):491–504
- Ebrahimi B, Tavana M, Kleine A, Dellnitz A (2021) An epsilon-based Data Envelopment Analysis approach for solving performance measurement problems with interval and ordinal dual-role factors. *OR Spectr* 43:1103–1124
- Emrouznejad A, Yang G-L (2018) A survey and analysis of the first 40 years of scholarly literature in DEA: 1978–2016. *Socioecon Plan Sci* 61:4–8
- Farrell MJ (1957) The measurement of productive efficiency. *J R Stat Soc Ser A (Gen)* 120(3):253–290
- Gouveia M, Dias L, Antunes C (2008) Additive DEA based on MCDA with imprecise information. *J Oper Res Soc* 59:54–63
- Gouveia M, Dias LC, Antunes CH (2013) Super-efficiency and stability intervals in additive DEA. *J Oper Res Soc* 64(1):86–96

- Greco S, Mousseau V, Słowiński R (2008) Ordinal regression revisited: multiple criteria ranking using a set of additive value functions. *Eur J Oper Res* 191(2):416–436
- Greco S, Ehrgott M, Figueira J (2016) Multiple criteria decision analysis—state of the art surveys. International series in operations research & management science. Springer, New York
- Hadi-Vencheh A, Matin RK (2011) An application of IDEA to wheat farming efficiency. *Agric Econ* 42(4):487–493
- Haghighat MS, Khorram E (2005) The maximum and minimum number of efficient units in DEA with interval data. *Appl Math Comput* 163(2):919–930
- Hosseinzadeh Lotfi F, Jahanshahloo GR, Khodabakhshi M, Rostamy-Malkhelifeh M, Moghaddas Z, Vaez-Ghasemi M (2013) A review of ranking models in Data Envelopment Analysis. *J Appl Math* 2013:492421
- Jahanshahloo GR, Lofti FH, Moradi M (2004) Sensitivity and stability analysis in DEA with interval data. *Appl Math Comput* 156(2):463–477
- Jiang B, Yang C, Dong Q, Li J (2021) Ecological efficiency evaluation of China's port industries with imprecise data. *Evol Intell*. <https://doi.org/10.1007/s12065-021-00638-2>
- Kadziński M, Tervonen T (2013) Robust multi-criteria ranking with additive value models and holistic pair-wise preference statements. *Eur J Oper Res* 228(1):169–180
- Kadziński M, Greco S, Słowiński R (2012) Extreme ranking analysis in robust ordinal regression. *Omega* 40(4):488–501
- Kadziński M, Labijak A, Napieraj M (2017) Integrated framework for robustness analysis using ratio-based efficiency model with application to evaluation of Polish airports. *Omega* 67:1–18
- Kao C (2006) Interval efficiency measures in Data Envelopment Analysis with imprecise data. *Eur J Oper Res* 174(2):1087–1099
- Kao C, Liu S-T (2005) Data Envelopment Analysis with imprecise data: an application of Taiwan machinery firms. *Int J Uncertain Fuzziness Knowl Based Syst* 13(2):225–240
- Kao C, Liu S-T (2009) Stochastic Data Envelopment Analysis in measuring the efficiency of Taiwan commercial banks. *Eur J Oper Res* 196(1):312–322
- Karsak E, Karadayi M (2017) Imprecise DEA framework for evaluating health-care performance of districts. *Kybernetes* 46(4):706–727
- Kim S-H, Park C-G, Park K-S (1999) An application of Data Envelopment Analysis in telephone offices evaluation with partial data. *Comput Oper Res* 26(1):59–72
- Labijak-Kowalska A, Kadziński M (2021) Experimental comparison of results provided by ranking methods in Data Envelopment Analysis. *Expert Syst Appl* 173:114739
- Labijak-Kowalska A, Kadziński M, Sychała I, Dias LC, Fiallos J, Patrick J, Michalowski W, Farion K (2023) Performance evaluation of emergency department physicians using robust value-based additive efficiency model. *Int Trans Oper Res* 30(1):503–544
- Lahdelma R, Salminen P (2001) SMAA-2: stochastic multicriteria acceptability analysis for group decision making. *Oper Res* 49(3):444–454
- Lahdelma R, Salminen P (2006) Stochastic multicriteria acceptability analysis using the data envelopment model. *Eur J Oper Res* 170(1):241–252
- Lahdelma R, Miettinen K, Salminen P (2003) Ordinal criteria in stochastic multicriteria acceptability analysis (SMAA). *Eur J Oper Res* 147(1):117–127
- Liang Q, Liao X, Shang J (2020) A multiple criteria approach integrating social ties to support purchase decision. *Comput Ind Eng* 147:106655
- Liu JS, Lu LY, Lu W-M, Lin BJ (2013) A survey of DEA applications. *Omega* 41(5):893–902
- Meyer P, Bigaret S (2012) Diviz: a software for modeling, processing and sharing algorithmic workflows in MCDA. *Intell Decis Technol* 6(4):283–296
- Park K (2007) Efficiency bounds and efficiency classifications in DEA with imprecise data. *J Oper Res Soc* 58(4):533–540
- Pelissari R, Oliveira MC, Abackerli AJ, Ben-Amor S, Assumpcao MRP (2021) Techniques to model uncertain input data of multi-criteria decision-making problems: a literature review. *Int Trans Oper Res* 28(2):523–559
- Saen RF (2006) Technologies ranking in the presence of both cardinal and ordinal data. *Appl Math Comput* 176(2):476–487
- Salo A, Punkka A (2011) Ranking intervals and dominance relations for ratio-based efficiency analysis. *Manag Sci* 57(1):200–214
- Sarkis J (2007) Preparing your data for DEA. In: Zhu J, Cook WD (eds) *Modeling data irregularities and structural complexities in Data Envelopment Analysis*. Springer, Berlin, pp 305–320

- Seiford L, Zhu J (1998) Stability regions for maintaining efficiency in Data Envelopment Analysis. *Eur J Oper Res* 108:127–139
- Shen Y, Hermans E, Brijs T, Wets G (2013) Data Envelopment Analysis for composite indicators: a multiple layer model. *Soc Indic Res* 114(2):739–756
- Shokouhi AH, Hatami-Marbini A, Tavana M, Saati S (2010) A robust optimization approach for imprecise Data Envelopment Analysis. *Comput Ind Eng* 59(3):387–397
- Tervonen T, Lahdelma R (2007) Implementing stochastic multicriteria acceptability analysis. *Eur J Oper Res* 178(2):500–513
- Tervonen T, van Valkenhoef G, Baştürk N, Postmus D (2013) Hit-And-Run enables efficient weight generation for simulation-based multiple criteria decision analysis. *Eur J Oper Res* 224(3):552–559
- Toloo M, Keshavarz E, Hatami-Marbini A (2021) An interval efficiency analysis with dual-role factors. *OR Spectr* 43:255–287
- Tomažević N, Seljak J, Aristovnik A (2016) TQM in public administration organisations: an application of Data Envelopment Analysis in the police service. *Total Qual Manag Bus Excell* 27(11–12):1396–1412
- Widiarto I, Emrouznejad A (2015) Social and financial efficiency of Islamic microfinance institutions: a Data Envelopment Analysis application. *Socioecon Plan Sci* 50:1–17
- Wu J, Yu Y, Zhu Q, An Q, Liang L (2018) Closest target for the orientation-free context-dependent DEA under variable returns to scale. *J Oper Res Soc* 69(11):1819–1833
- Zahran SZ, Alam JB, Al-Zahrani AH, Smirlis Y, Papadimitriou S, Tsioumas V (2020) Analysis of port efficiency using imprecise and incomplete data. *Oper Res Int J* 20(1):219–246
- Zhu J (1996) Robustness of the efficient DMUs in Data Envelopment Analysis. *Eur J Oper Res* 90:451–460
- Zhu J (2003) Imprecise Data Envelopment Analysis (IDEA): a review and improvement with an application. *Eur J Oper Res* 144(3):513–529

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Publication [P6]

A. Labijak-Kowalska, M. Kadziński, and L. C. Dias. Robustness analysis for imprecise additive value efficiency analysis with an application to evaluation of special economic zones in poland. 2023. Submitted to Socio-Economic Planning Sciences.

Contribution of co-authors (excluding the author and the supervisor of this dissertation):

- Luis C. Dias
 - Co-authorship of the idea underlying the paper consisting in conducting robustness analysis in the context of value-based efficiency analysis,
 - consultations on the proposed value-based concepts and respective mathematical models ,
 - review and editing of the manuscript.

Robustness Analysis for Imprecise Additive Value Efficiency Analysis with an Application to Evaluation of Special Economic Zones in Poland

Anna Labijak-Kowalska^a, Miłosz Kadziński^{a,*}, Luis C. Dias^b

^a*Poznan University of Technology, Faculty of Computing and Telecommunications, Piotrowo 2, 60-965 Poznań, Poland*

^b*University of Coimbra, CeBER, Faculty of Economics, Av. Dias da Silva n.165, Coimbra, 3004-512, Portugal*

Abstract

We introduce an algorithmic framework for investigating the robustness of efficiency analysis results in the presence of imprecise information about data and preferences. We employ an additive value efficiency model accepting ordinal and interval information about performances of Decision Making Units and imprecision in the specification of input and output weights and the shapes of marginal value functions. We verify the stability of efficiency measures using a combination of mathematical programming and Monte Carlo simulations. The results capture various certainty levels, emphasizing the necessary, possible, extreme, and expected outcomes and the distribution of outcomes in the space of feasible weights, performances, and marginal functions. The practical usefulness of the proposed framework is demonstrated in a real-world problem concerning the functioning of Special Economic Zones in Poland. We discuss results that increase the discrimination power, indicate overall good performances, and provide hints on the required improvements.

Keywords: Value-based efficiency analysis, Robustness analysis, Imprecision, Special economic zone, Open-source software in R, Multiple criteria decision analysis

1. Introduction

Operations Research (OR) deals with applying analytical tools, methods, and techniques to study complex decision problems. A general aim of OR consists of providing managers with a sound basis for decision-making, i.e., recommendations on the solution of problems concerning a given system's operations. The crucial steps of the OR process involve identifying a problem to be solved, constructing a mathematical model that adequately reflects the real-world situation with its variables, objectives, and constraints, and employing the model for deriving the solutions.

OR has been extensively used in industry, business, non-profit organizations, and government, supporting the introduction of operational improvements. Under the umbrella of OR, a broad range of problem-solving sub-areas has been developed, with Data Envelopment Analysis (DEA) and Multiple Criteria Decision Analysis (MCDA) being among the most popular streams. As noted in [5], the primary aims of DEA and MCDA are different, and the two fields were developed, to a large extent, independently of each other. On the one hand, DEA is a non-parametric method for estimating a best-practice frontier

*Corresponding author: Institute of Computing Science, Poznań University of Technology, Piotrowo 2, 60-965 Poznań, Poland. Tel. +48-61 665 3022.

Email addresses: anna.labijak@cs.put.poznan.pl (Anna Labijak-Kowalska), milosz.kadziński@cs.put.poznan.pl (Miłosz Kadziński), lmcdias@fe.uc.pt (Luis C. Dias)

and measuring the relative efficiency of Decision Making Units (DMUs) in converting inputs into outputs. In this regard, DEA has been focussed on analyzing historical, objective data to control, monitor, and assess the performance of DMUs within the ex-post evaluation framework. On the other hand, MCDA offers a diversity of approaches that involve the Decision Makers (DMs) in carrying forward the solution of ranking, choice, or sorting problems involving multiple, potentially conflicting criteria, reflecting different viewpoints on the quality of decision alternatives. Hence, it has been oriented toward the ex-ante evaluation of alternatives, incorporating subjectivity in the form of DMs' preferences.

Even though the background and philosophy underlying DEA and MCDA differ, one can also indicate significant similarities between the two fields. In particular, they coincide when considering DMUs as alternatives, inputs as cost criteria (costs), and outputs as gain criteria (benefits) [6]. Moreover, the models used to measure the efficiency of DMUs in DEA and a distance to the Pareto frontier in MCDA are equivalent or strongly related [2, 13, 27]. In this perspective, the concepts of efficiency and Pareto optimality build on the same idea of attaining the most advantageous score for at least one feasible scenario or set of parameters. These similarities have implied an ever-growing overlap between the two fields in terms of problems faced and concepts incorporated for decision support and problem structuring. In recent years, the two fields have learned from each other in what concerns efficiency or preference models, ways of structuring decision problems, incorporated value judgments or preference information, aims of the performed analysis, and applied solution methods. Thus, they are more often perceived as complementary rather than competing methods.

When it comes to the employed models, the traditional way of quantifying efficiency in DEA involves the ratio of virtual outputs and inputs, i.e., weighted sums of outputs and inputs, respectively. In MCDA, various families of preference models are used to aggregate the alternatives' performances on multiple criteria in line with the DMs' preferences. The major families of such models – offering intuitiveness and high explanatory capabilities – include scoring functions, binary relations, and decision rules. Recently various scoring functions, initially used in MCDA, have been incorporated into DEA methods. The two examples of such models include a value function and the Choquet integral [21]. On the one hand, [11] proposed to convert inputs and outputs into value functions aggregated using an additive value model and select for each DMU the weights minimizing the value difference to the best DMU. On the other hand, [22] employed the Choquet multiple criteria preference aggregation model to account for interactions between different inputs or outputs.

As far as structuring the decision problem is concerned, many MCDA methods have incorporated a hierarchical structure of criteria [3]. It proved to be efficient from both structural and computational viewpoints, allowing decomposing the considered problems into manageable pieces and analyzing results at different levels and nodes of the hierarchy. This idea has been recently incorporated into the DEA models (see, e.g., [26]). Multiple layer models allow for a more detailed analysis, greater discriminating power, and more realistic weight allocations.

When it comes to measuring the performance of DMUs, DEA models initially did not involve subjective information about the importance of different inputs and outputs. On the contrary, incorporating value judgments into the model and deriving recommendations consistent with the DM's preference information has always been at the core of MCDA. To constrain the flexibility of weights and prevent some units from being judged as efficient based on unrealistically extreme weights (e.g., when only a single input or a single output is assigned a non-zero weight), subjective weight restrictions have also been incorporated in DEA.

These are usually imprecise, taking the form of, e.g., absolute bounds on a single weight value, admissible intervals for the ratio between two weights, or linear weight constraints. They may be motivated by the analysis of market prices, expert opinion, or preferences of the DM controlling the units. For a detailed description of the role of preference information and various types of judgments applicable in the context of DEA, see [2, 16]. In fact, some MCDA tools have been explicitly used for restricting weight values of a DEA model [4].

Apart from investigating the efficiency status, DEA determines the benchmarks for the inefficient DMUs. In this way, the latter are provided with recommendations on reducing the inputs or increasing the outputs and, in general, whom to follow to become efficient. MCDA has been traditionally focused on deriving the outcomes for a decision problem while being less concerned with improving the recommendation for a given alternative. In this regard, sensitivity analysis has been oriented toward investigating if the outcomes would change if we applied some other parameter values of the preference model. However, a number of approaches have been recently proposed to indicate the performance modifications allowing the attainment of a particular decision target [15, 17, 23]. These methods either minimize the changes required on different criteria or balance various steps within a stepwise benchmarking framework.

The primary aim of DEA is to divide the DMUs into two groups: efficient and inefficient ones. In many decision problems, such binary classification is insufficient due to its weak discrimination power. This is particularly evident for problems involving many DMUs and numerous inputs and outputs when the units become too specialized. Moreover, the traditional DEA models evaluate each DMU in terms of the weights, which are the most advantageous for it. Motivated by the MCDA advancements, two different methodological streams have been developed in DEA. On the one hand, several ranking approaches have been proposed to order the units from the best to the worst [1, 14]. Some of these methods are based on a common set of weights, offering a joint criterion for the evaluation of DMUs. On the other hand, robustness analysis methods have been used to investigate the variability of efficiency outcomes for all feasible input and output weights [18, 25].

This paper contributes to the cross-fertilization of MCDA and DEA. Specifically, we propose a novel framework for efficiency analysis that derives from the recent developments in MCDA and creatively incorporates them into a DEA-based method. The following three features distinguish the introduced framework.

First, we build on the Value-based Additive DEA [11]. The employed efficiency model converts inputs and outputs into monotonic criteria [9]. All criteria are associated with marginal value functions, which are, in turn, aggregated using an additive function into a comprehensive score. In MCDA, such a model is popular due to a straightforward interpretation of numerical scores and the possibility of meaningful aggregation of performances expressed on different scales. Despite these favorable features, its application and development in the context of DEA are still in their infancy compared to, e.g., ratio-based models.

Second, we admit imprecision in the efficiency analysis [28]. On the one hand, we account for the imprecise evaluations in terms of intervals of possible values or ordinal assessments. Such imprecision and ill-determination are common in real-world decision problems [10]. Thus we tolerate uncertainty and availability of qualitative information in addition to precise quantitative measurements. On the other hand, we consider imprecision in the efficiency model. In particular, the marginal value functions need to fit in the space delimited by pre-defined extreme shapes, whereas linear inequalities constrain the variability of input/output weights. In this way, we can handle different curvatures for the individual factors and their

various impacts on the comprehensive efficiency measure, e.g., with a convex curve indicating that the same amount of variation is more valued for higher input or output levels, and a concave curve indicating the opposite.

Third, we propose tools for analyzing the robustness of four types of efficiency results: distance to the best DMU, scores, preference relations, and ranks [19]. These tools quantify the variability of outcomes that can be attained for all feasible scenarios founded by the imprecision of input/output performances, marginal value functions, and weights. For each perspective, we compute extreme, robust outcomes univocally confirmed by all scenarios, and also stochastic indices, quantifying the probability of a given result in the space of feasible scenarios. They are computed with mathematical programming and Monte Carlo simulations, respectively. The proposed methods for robustness analysis are made available in the form of open-source models implemented in R. Hence the user can adjust the outcomes that are of interest for a given problem.

We illustrate the introduced framework and software in a real-world study concerning the efficiency of Special Economic Zones (SEZs) in Poland. SEZ is a designated area where businesses can be run under preferential conditions. We consider 14 zones that are described in terms of two inputs, including total area and capital expenditures, and two outputs, including the number of jobs and financial results. The area and the number of jobs are considered interval factors delimited by the extreme values observed in the analyzed term. For each input and output, we account for an acceptable range of marginal value functions. The discussed results demonstrate the practical usefulness of robustness analysis outcomes.

The paper's remainder is organized in the following way. Section 2 reminds Value-based Additive DEA and introduces types of imprecision handled by the proposed framework. In Section 3, we discuss the robust outcomes computed with mathematical programming and Monte Carlo simulation. Section 4 discusses the open-source software implementation in R. In Section 5, we demonstrate the results and benefits of using robustness analysis in the context of real-world data concerning Polish SEZs. The last section concludes the paper and outlines avenues for future research.

2. Dealing with Imprecision in Value-based Additive Efficiency Analysis

Let us use the following notation:

- \mathcal{D} – a finite set of K DMUs, $\mathcal{D} = \{DMU_1, \dots, DMU_K\}$,
- I and O – the set of M inputs and N outputs, respectively,
- $x_{q,k}$ – the performance of DMU_k on input $q \in I$, and $y_{q,k}$ – the performance of DMU_k on output $q \in O$,
- $Q = N + M$ – a number of all factors relevant for the analysis,
- w_q – a weight associated with the q -th factor (input or output); the weights are normalized to sum up to one, i.e., $\sum_{q=1}^Q w_q = 1$,
- u_q – a marginal value function associated with the q -th factor,
- $S_w = \{w = (w_1, w_2, \dots, w_q)^T | w \geq 0, A_w w \leq 0\}$ – a space of feasible weight vectors, where A_w is the coefficient matrix of user-defined linear weight constrains.

In what follows, we first remind the Value-based Additive Data Envelopment Analysis (VDEA). Then, we introduce types of handled imprecise data related to the specification of factors' weights, units' performances, and marginal value functions. Finally, we discuss handling such imprecision in exact methods based on mathematical programming and stochastic approaches incorporating the Monte Carlo simulations.

2.1. Reminder on Value-based Additive Efficiency Analysis

DEA incorporates various models to quantify the efficiency of DMUs. The most common is the ratio-based model, where the efficiency is expressed as the ratio between the weighted sum of outputs and the weighted sum of inputs [7]. However, in recent years, other models have also found use in practical studies. In particular, [11] proposed a value-based additive efficiency measure inspired by Multi-Attribute Value Theory (MAVT) and the additive DEA model. It associates a monotonic marginal value function u_q with each factor q . The function takes values in the range $[0, 1]$, being non-increasing for inputs $q \in I$ (and for any undesirable outputs) and non-decreasing for (desirable) outputs $q \in O$. Then, the comprehensive value of DMU_o is defined using an additive value function:

$$E_o = \sum_{q=1}^Q w_q u_q(DMU_o). \quad (1)$$

To verify the efficiency of DMU_o , one needs to minimize its distance to the unit with the greatest comprehensive value:

Minimize d_o

$$\text{s.t.} \quad \left. \begin{array}{l} \sum_{q=1}^Q w_q u_q(DMU_k) - \sum_{q=1}^Q w_q u_q(DMU_o) \leq d_o, \text{ for } k = 1, \dots, K, \\ d_o \geq 0, \\ \sum_{q=1}^Q w_q = 1, \\ w_q \geq 0, \text{ } q = 1, \dots, Q, \\ \mathbf{w} \in S_w. \end{array} \right\} \mathcal{W} \quad (2)$$

In the spirit of DEA, this formulation allows DMU_o to select the weights that place it as close as possible to the frontier of efficiency defined by the set of DMUs, the value functions and the weight constraints. If the optimal value $d_{*,o}$ equals zero, DMU_o is judged efficient, attaining the greatest comprehensive value among the considered DMUs for at least one feasible weight vector. If $d_{*,o}$ is greater than zero, DMU_o is inefficient.

2.2. Imprecise Performances and Preferences in Value-based Efficiency Analysis

In the standard VDEA, inputs and outputs are deterministic and the marginal value functions are precisely specified. However, in real-world decision problems, the performances of the DMUs in some factors (inputs or outputs) may be uncertain. We consider two types of uncertainty: the performances can be given in the form of intervals [12] (which might reflect imprecision, hesitation, or variation), or the performances can be given as a ranking (ordinal performances) where neither the differences nor ratios are known. Moreover, the users may wish to account for ranges of admissible values for the cardinal factors instead of pre-defined exact shapes of marginal values. To incorporate such imprecisions into the efficiency model, let us consider the following notation:

- II and IO – the sets of imprecise inputs and outputs; $II \subseteq I$ and $IO \subseteq O$,
- OI and OO – the sets of ordinal inputs and outputs; $OI \subseteq I$ and $OO \subseteq O$,
- $[x_{q,k*}, x_{q,k}^*]$ – the interval of possible input values for $q \in II$ for DMU_k ,
- $[y_{q,k*}, y_{q,k}^*]$ – the interval of possible output values for $q \in IO$ for DMU_k ,
- $u_{q,*}(x_{q,k})$ – the lower marginal value assigned to input performance $x_{q,k}$, and $u_{q,*}(y_{q,k})$ – the lower marginal value assigned to output performance $y_{q,k}$,
- $u_q^*(x_{q,k})$ – the upper marginal value assigned to input performance $x_{q,k}$, and $u_q^*(y_{q,k})$ – the upper marginal value assigned to output performance $y_{q,k}$,
- $U_{q,k}$ – the product of weight w_q and the performance of DMU_k on factor q , i.e., $w_q \cdot u_q(x_{q,k})$ or $w_q \cdot u_q(y_{q,k})$.

Example: Let us illustrate the proposed methods while referring to the efficiency analysis of twelve hospitals described in terms of four inputs (i_1 – i_4) and two outputs (o_1 – o_2). For the sake of this illustration, the data is modified after [8] by introducing imprecision into the problem definition. We consider three interval inputs, one ordinal input, and two interval outputs. The inputs are i_1 – the number of doctors (interval), i_2 – doctors’ cost (ordinal), i_3 – the number of nurses (interval), and i_4 – nurses’ cost (interval). The outputs are o_1 – the number of inpatients (interval) and o_2 – the number of outpatients (interval). The performances are provided in Table 1. For all imprecise factors, we consider extreme marginal value functions delimiting the range of admissible values (see Figure 1). In turn, for i_2 , only the order of performances matters. Note that the first position means the lowest (best) input value in this case. Also, we restrict the weight of each factor to the range $[0.083, 0.250]$, hence preventing it from being dominant or negligible. Thus, the constraint set \mathcal{W} for this example ensures that $w_q \geq 0.083$ and $w_q \leq 0.250$ for each $q = 1, 2, \dots, 6$, and $\sum_{q=1}^6 w_q = 1$.

Table 1: Input and output performances for the illustrative example concerning twelve hospitals.

DMU	i_1	i_2	i_3	i_4	o_1	o_2
H_1	[24, 24]	8th	[154, 161]	[98, 100]	[90, 95]	[85, 88]
H_2	[17, 19]	2nd	[124, 131]	[72, 76]	[170, 182]	[80, 85]
H_3	[23, 25]	7th	[142, 150]	[85, 90]	[172, 180]	[60, 63]
H_4	[45, 51]	9th	[170, 178]	[135, 148]	[120, 140]	[48, 50]
H_5	[15, 17]	1st	[147, 155]	[58, 62]	[96, 102]	[69, 73]
H_6	[60, 65]	7th	[252, 255]	[85, 95]	[218, 255]	[85, 90]
H_7	[35, 42]	8th	[232, 235]	[98, 100]	[190, 200]	[83, 88]
H_8	[31, 31]	7th	[205, 206]	[85, 85]	[130, 140]	[72, 75]
H_9	[28, 30]	3rd	[231, 244]	[72, 76]	[195, 215]	[105, 110]
H_{10}	[47, 50]	5th	[258, 268]	[72, 75]	[240, 250]	[97, 100]
H_{11}	[50, 53]	6th	[301, 306]	[78, 80]	[280, 292]	[142, 147]
H_{12}	[35, 38]	4th	[213, 250]	[60, 65]	[250, 255]	[113, 120]

2.3. Dealing with Imprecision in Mathematical Programming

In this section, we discuss how imprecise performances and preferences are treated in mathematical programming models.

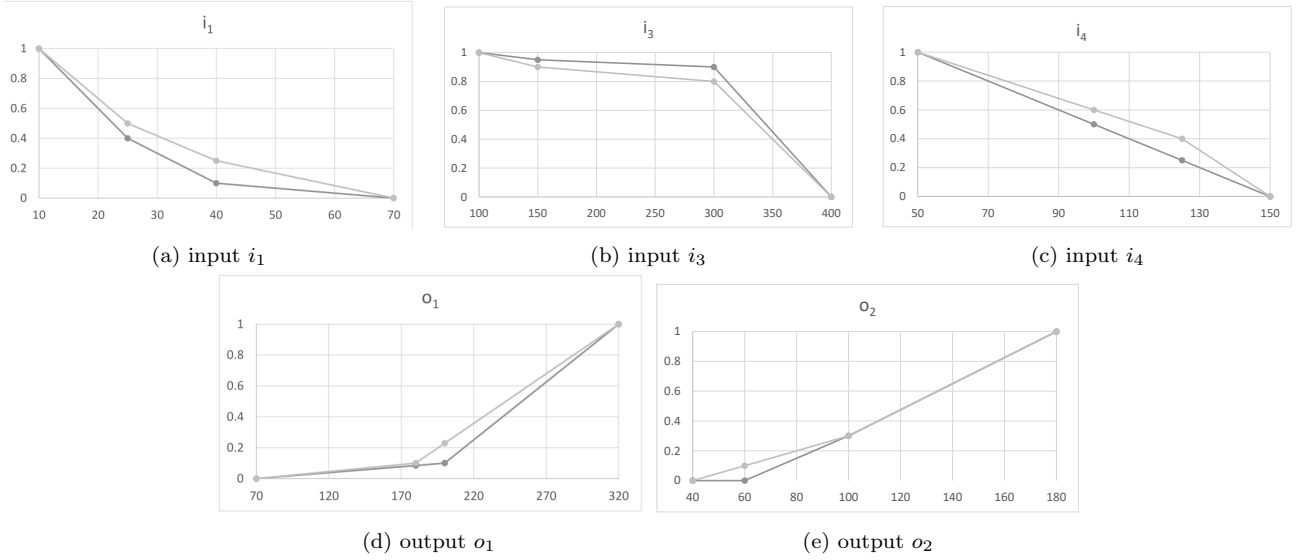


Figure 1: The extreme marginal values functions for the interval inputs and outputs for illustrative example.

Interval factors. The interval performances are replaced with the precise ones depending on the considered type of result. When identifying the best possible result for DMU_o , we use the most favorable (optimistic) scenario for this unit and the least favorable scenarios for the remaining ones ($SCE = OPT$), i.e.:

$$x_{q,k}^{OPT,o} = \begin{cases} x_{q,k*}, & \text{if } k = o, \\ x_{q,k}^*, & \text{otherwise,} \end{cases} \quad (3)$$

$$y_{q,k}^{OPT,o} = \begin{cases} y_{q,k}^*, & \text{if } k = o, \\ y_{q,k*}, & \text{otherwise.} \end{cases} \quad (4)$$

For example the most favorable scenario for H_1 involves the minimal input performances: $x_{1,1} = 24$, $x_{3,1} = 154$, and $x_{4,1} = 98$ and the maximal output performances: $y_{1,1} = 95$ and $y_{2,1} = 88$. For all other units, we take the maximal inputs (e.g., $x_{1,2} = 19$, $x_{3,7} = 235$) and minimal outputs (e.g., $y_{1,2} = 170$ and $y_{2,5} = 69$). Let $\pi_{q,o}^{OPT}$ denote the permutation of units that reorders them according to a non-decreasing order of their performances ($x_{q,k}^{OPT,o}$ or $y_{q,k}^{OPT,o}$, $k = 1, \dots, K$) on factor q in the optimistic scenario for unit DMU_o . Hence $\pi_{q,o}^{OPT}(k)$ is the index of the unit with the k -th least performance on factor q when considering the performances in the optimistic scenario for DMU_o . This will be needed to ensure that the accepted performance variation within the intervals does not contradict this ranking.

When looking for the worst possible outcomes for DMU_o , we replace the intervals with the least favorable (pessimistic) scenario for this unit and the most favorable scenarios for the remaining ones ($SCE = PES$), i.e.:

$$x_{q,k}^{PES,o} = \begin{cases} x_{q,k}^*, & \text{if } k = o, \\ x_{q,k*}, & \text{otherwise,} \end{cases} \quad (5)$$

$$y_{q,k}^{PES,o} = \begin{cases} y_{q,k*}, & \text{if } k = o, \\ y_{q,k}^*, & \text{otherwise.} \end{cases} \quad (6)$$

In the context of H_1 from the considered example, we would use the maximal input performances ($x_{1,1} = 24$,

$x_{3,1} = 161$, $x_{4,1} = 100$) and the minimal output performances ($y_{1,1} = 90$ and $y_{2,1} = 85$). For all other units, we take the minimal inputs (e.g., $x_{1,2} = 17$ and $x_{3,7} = 232$) and the maximal outputs (e.g., $y_{1,2} = 182$ and $y_{2,5} = 73$). Let $\pi_{q,o}^{PES}$ denote the permutation of units that reorders them according to a non-decreasing order of their performances ($x_{q,k}^{PES,o}$ or $y_{q,k}^{PES,o}$, $k = 1, \dots, K$) on factor q in the pessimistic scenario for unit DMU_o . Hence $\pi_{q,o}^{PES}(k)$ is the index of the unit with the k -th least performance on factor q when considering the performance in the pessimistic scenario for DMU_o .

Ordinal factors. For ordinal inputs or outputs, we reproduce the order of performances. For this purpose, we replace the product $w_q \cdot u_q(DMU_k)$ with a single variable $U_{q,k}$. Let π_q denote the permutation of units that reorders them according to a non-decreasing order of their performances on factor q , i.e., $x_{q,\pi_q(1)}, \dots, x_{q,\pi_q(k)}, x_{q,\pi_q(k+1)}, \dots, x_{q,\pi_q(K)}$. Then, if $x_{q,\pi_q(k)}$ and $x_{q,\pi_q(k+1)}$ are the same, we set $U_{q,\pi_q(k)} = U_{q,\pi_q(k+1)}$. On the contrary, if $x_{q,\pi_q(k+1)} > x_{q,\pi_q(k)}$, then in the case $q \in IO$, we set $U_{q,\pi_q(k)} \geq \alpha U_{q,\pi_q(k+1)}$, and when $q \in OO$, we set $\alpha U_{q,\pi_q(k)} \leq U_{q,\pi_q(k+1)}$, for some $\alpha > 1$. In addition, we ensure that the values assigned to the worst performances are positive by comparing them with an arbitrarily small positive constant ϵ , whereas the values associated with the best performances are not greater than the respective weight w_q . The set of constraints for modeling the ordinal factors is as follows:

$$\begin{array}{l}
\text{for } q \in OI : \\
\left. \begin{array}{l}
U_{q,\pi_q(K)} \geq \epsilon, \quad U_{q,\pi_q(1)} \leq w_q, \\
U_{q,\pi_q(k)} \geq \alpha U_{q,\pi_q(k+1)}, \quad \text{for } k = 1, \dots, K-1, \text{ such that } x_{q,\pi_q(k+1)} > x_{q,\pi_q(k)}, \\
U_{q,\pi_q(k)} = U_{q,\pi_q(k+1)}, \quad \text{for } k = 1, \dots, K-1, \text{ such that } x_{q,\pi_q(k+1)} = x_{q,\pi_q(k)}
\end{array} \right\} \left. \begin{array}{l}
ORD^{OI}(q, \pi_q) \\
ORD(OI, OO)
\end{array} \right\} \\
\text{for } q \in OO : \\
\left. \begin{array}{l}
U_{q,\pi_q(1)} \geq \epsilon, \quad U_{q,\pi_q(K)} \leq w_q, \\
\alpha U_{q,\pi_q(k)} \leq U_{q,\pi_q(k+1)}, \quad \text{for } k = 1, \dots, K-1, \text{ such that } y_{q,\pi_q(k+1)} > y_{q,\pi_q(k)}, \\
U_{q,\pi_q(k)} = U_{q,\pi_q(k+1)}, \quad \text{for } k = 1, \dots, K-1, \text{ such that } y_{q,\pi_q(k+1)} = y_{q,\pi_q(k)}.
\end{array} \right\} \left. \begin{array}{l}
ORD^{OO}(q, \pi_q) \\
ORD(OI, OO)
\end{array} \right\}
\end{array}$$

Note that to enforce monotonicity, it would also be possible to incorporate an additive form of the constraint (e.g., $U_{q,\pi_q(i)} + \epsilon \leq U_{q,\pi_q(i+1)}$ when $q \in OO$). Here, we opt for the multiplicative representation as typically considered in DEA [28]. In the considered example, for input i_2 , the least performance is attained by H_5 , followed by H_2 and H_9 , and the greatest input is consumed by H_4 . This implies the following constraints $w_2 \geq U_{2,5}$, $U_{2,5} \geq \alpha U_{2,2}$, $U_{2,2} \geq \alpha U_{2,9}$, \dots, \dots , $U_{2,1} \geq \alpha U_{2,4}$, $U_{2,4} \geq \epsilon$.

Value functions ranges. Having replaced the interval input (output) performances with the precise ones, we need to ensure that the marginal values $u_q(x_{q,k})$ ($u_q(y_{q,k})$) are between the extreme admissible regions, i.e., $U_{q,k} \geq w_q u_{q,*}(x_{q,k})$ and $U_{q,k} \leq w_q u_q^*(x_{q,k})$ ($U_{q,k} \geq w_q u_{q,*}(y_{q,k})$ and $U_{q,k} \leq w_q u_q^*(y_{q,k})$). Moreover, we impose the monotonicity constraints for the marginal values assigned to the performances observed for all DMUs, similarly as for the ordinal factors. Overall, the constraint set depends on whether we consider the

optimistic ($\pi_{q,o}^{SCE}(k) = \pi_{q,o}^{OPT}(k)$) or pessimistic ($\pi_{q,o}^{SCE}(k) = \pi_{q,o}^{PES}(k)$) scenario for DMU_o :

$$\left. \begin{array}{l}
\text{for } q \in II : \\
U_{q,\pi_{q,o}^{SCE}(k)} \geq w_q u_{q,*}(x_{q,\pi_{q,o}^{SCE}(k)}), \quad k = 1, \dots, K, \\
U_{q,\pi_{q,o}^{SCE}(k)} \leq w_q u_q^*(x_{q,\pi_{q,o}^{SCE}(k)}), \quad k = 1, \dots, K, \\
ORD^{OI}(q, \pi_{q,o}^{SCE}(k)), \\
\text{for } q \in IO : \\
U_{q,\pi_{q,o}^{SCE}(k)} \geq w_q u_{q,*}(y_{q,\pi_{q,o}^{SCE}(k)}), \quad k = 1, \dots, K, \\
U_{q,\pi_{q,o}^{SCE}(k)} \leq w_q u_q^*(y_{q,\pi_{q,o}^{SCE}(k)}), \quad k = 1, \dots, K, \\
ORD^{OO}(q, \pi_{q,o}^{SCE}(k)).
\end{array} \right\} \begin{array}{l}
IMP^{II}(q, \pi_{q,o}^{SCE}) \\
IMP^{IO}(q, \pi_{q,o}^{SCE})
\end{array} \left. \vphantom{\begin{array}{l} \text{for } q \in II : \\ \text{for } q \in IO : \end{array}} \right\} IMP^{SCE}(II, IO) \quad (8)$$

For example, when considering interval input i_1 and hospital H_1 ($x_{1,1} = 24$), the respective marginal value must be between the lower and upper admissible values, $u_{1,*}(24) \leq u_1(24) \leq u_1^*(24)$. Given the extreme shapes of marginal value functions, the respective constraint takes the following form: $0.533 \leq u_1(24) \leq 0.627$. When considering the substitute variable $U_{1,1}$ corresponding to the minimal performance (24) on the first factor (i_1), it translates into: $0.533 \cdot w_1 \leq U_{1,1} \leq 0.627 \cdot w_1$.

When it comes to output o_1 and considering $DMU_o = H_1$, in the optimistic scenario for H_1 , the performances are: 95, 170, 172, 120, 96, 218, 190, 130, 195, 240, 280, and 250. The corresponding permutation π_{q,H_1}^{OPT} of units in the non-decreasing order of performances is $H_1, H_5, H_4, H_8, H_2, H_3, H_7, H_9, H_6, H_{10}, H_{12}$, and H_{11} . The monotonicity constraints, in this case, take the following form: $U_{5,1} \geq \epsilon$, $\alpha U_{5,1} \leq U_{5,5}$, $\alpha U_{5,5} \leq U_{5,4}$, \dots , $\alpha U_{5,12} \leq U_{5,11}$, and $U_{5,11} \leq w_5$.

2.4. Dealing with Imprecision in Stochastic Methods

Another way to conduct the robustness analysis consists of sampling the space of feasible weights, performances, and marginal value functions using Monte Carlo simulation. To conduct the sampling, we implement a three-stage procedure based on the Hit-And-Run algorithm. In the first step, we generate the pre-defined number (T) of weight vectors $w^{(t)}$ from space \mathcal{W} . Hence they are non-negative, normalized to sum to one, and satisfy the constraints specified by users. In the second step, we draw the performance for the interval inputs and outputs from ranges $[x_{q,o*}, x_{q,o}^*]$ and $[y_{q,o*}, y_{q,o}^*]$. For example, in the illustrative study, the performance $x_{3,2}^{(t)}$ of hospital H_2 for input i_3 satisfies the following conditions: $x_{3,2}^{(t)} \geq 124$ and $x_{3,2}^{(t)} \leq 131$. In the third step, we sample the marginal values for all factors and units. For the ordinal factors, the marginal values need to be in the interval $[0, 1]$ and adhere to the monotonicity constraints. For example, in the illustrative study, the marginal values for i_2 need to be consistent with the following conditions: $u_{2,4}^{(t)} \geq \epsilon$, $\alpha u_{2,4}^{(t)} \leq u_{2,1}^{(i)}$, \dots , $u_{2,3}^{(t)} = u_{2,6}^{(t)}$, \dots , $u_{2,5}^{(t)} \leq 1$. For the interval factors with pre-defined admissible ranges of marginal functions, we take the precise performances from the previous point as the starting point and then draw the respective marginal values from the allowed intervals. For example, when considering the input performance $x_{q,o}^{(t)}$, the marginal value $u_{q,o}^{(t)}$ needs to satisfy the following conditions: $u_{q,*}(x_{q,o}^{(t)}) \leq u_{q,o}^{(t)} \leq u_q^*(x_{q,o}^{(t)})$. Moreover, we ensure that the generated marginal values are monotonic. For example, if the performance $x_{3,2}^{(t)}$ drawn for H_2 for input i_3 was 130, then the marginal value needs to be between 0.94 ($u_{3,*}(130)$) and 0.97 ($u_3^*(130)$). Each derived sample t can be used to compute the comprehensive value for each DMU_o , which can be interpreted as an absolute efficiency score in relation to a

perfect DMU having a value of 1 on all value functions:

$$E_o^{(t)} = \sum_{q=1}^Q w_q^{(t)} u_{q,o}^{(t)}. \quad (9)$$

Given T uniformly distributed samples, we can summarize the distribution of various efficiency results. They fill the gap between the extreme outcomes or the necessary and the possible results.

3. Robustness Analysis for Imprecise Value-based Additive Efficiency Analysis

The proposed framework for robustness analysis comprises two types of methods, exact and stochastic. The exact methods focus on the extreme results, whereas the stochastic approaches quantify the probability of results attainable in the space of feasible weights, performances, and value functions. We consider four viewpoints related to the distances from the best DMU, absolute scores, ranks, and preference relations. The proposed methods are illustrated with a numerical example considering twelve hospitals.

3.1. Distance to the Best Decision Making Unit

In this section, we focus on the distance d_o of each DMU_o to the best unit. Note that such a distance corresponds to a DEA relative efficiency, i.e., relative to the empirically observed efficient frontier, which could change if the set of DMUs is modified. To find the best (minimal) distance, the following mathematical programming model must be solved:

$$\begin{aligned} & \text{Minimize } d_o \\ \text{s.t. } & \left. \begin{aligned} & \sum_{q=1}^Q U_{q,k} - \sum_{q=1}^Q U_{q,o} \leq d_o, \text{ for } k = 1, \dots, K, \\ & d_o \geq 0, \\ & \mathcal{W}, \text{ ORD}(OI, OO), \text{ IMP}^{OPT}(II, IO). \end{aligned} \right\} \quad (10) \end{aligned}$$

This model minimizes distance d_o , indicating the maximal difference between comprehensive values of any DMU_k , $k = 1, \dots, K$ and DMU_o subject to the constraints defining the set of feasible weights (\mathcal{W}) and maintaining the specificity of the ordinal ($\text{ORD}(OI, OO)$) and imprecise factors while assuming the optimistic scenario for DMU_o ($\text{IMP}^{OPT}(II, IO)$).

The previous efficiency score is generous for each DMU, as it allows the DMU to be evaluated under the most favorable weights, performance values, and value functions (subject to the defined constraints). On the opposite, one can search for the most disadvantageous relative assessment for the DMU, with the least favorable performances and value functions, and with the weighting vector that penalizes it the most. This amounts to finding the maximal distance of DMU_o to the best DMU, solving the following Mixed-Integer Linear Programming (MILP) model:

$$\begin{aligned} & \text{Maximize } d_o \\ \text{s.t. } & \left. \begin{aligned} & \sum_{q=1}^Q U_{q,k} - \sum_{q=1}^Q U_{q,o} \geq d_o - C(1 - b_k), \text{ for } k = 1, \dots, K, \\ & \sum_{k=1}^K b_k \geq 1, \\ & b_k \in \{0, 1\} \text{ for } k = 1, \dots, K, \\ & d_o \geq 0, \\ & \mathcal{W}, \text{ ORD}(OI, OO), \text{ IMP}^{PES}(II, IO), \end{aligned} \right\} \quad (11) \end{aligned}$$

where C is a large positive constant ($C \gg 1$). In this case, we maximize the distance d_o with the restriction that for at least one unit DMU_k , it is equal to the difference between E_k and E_o . To satisfy this condition, the binary variables $b_k \in \{0, 1\}$, $k = 1, \dots, K$, are introduced. If b_k is equal to 0, then $C(1 - b_k)$ is equal to C , and the constraint is always satisfied. In turn, when b_k is equal to 1, $C(1 - b_k)$ is equal to 0 and the constraint guarantees that d_o is equal to $E_k - E_o$. The remaining constraints define the set of feasible weights (\mathcal{W}) and maintain the specificity of the ordinal ($ORD(OI, OO)$) and imprecise factors while assuming the pessimistic scenario for DMU_o ($IMP^{PES}(II, IO)$).

Example. For the considered illustrative study, the extreme distances of twelve hospitals to the best one are presented in Table 2. There are five hospitals (H_2, H_5, H_9, H_{11} , and H_{12}) with the minimal distance (d_{o*}) equal to 0. They are deemed efficient. According to the minimal distance, the worst hospital is H_4 ($d_{4*} = 0.129$). When it comes to the worst (maximal) possible distance to the best hospital, H_5 attains the most favorable score ($d_2^* = 0.306$), being followed by H_2 ($d_2^* = 0.343$). Hospitals H_3 (0.608) and H_4 (0.702) are at the ranking's bottom in this regard.

Table 2: Extreme distances to the best hospital for the twelve hospitals in the illustrative study.

	H_1	H_2	H_3	H_4	H_5	H_6	H_7	H_8	H_9	H_{10}	H_{11}	H_{12}
$d_{o,*}$	0.016	0.000	0.012	0.129	0.000	0.015	0.035	0.041	0.000	0.012	0.000	0.000
d_o^*	0.582	0.343	0.608	0.702	0.306	0.558	0.590	0.600	0.422	0.459	0.422	0.403

When conducting the stochastic analysis, two distance-related outcomes are of particular interest:

- *Distance Acceptability Interval Index* ($DAII(DMU_o, b_i)$) is the share of feasible scenarios for which the distance of DMU_o to the best unit is in the interval $b_i \subseteq [0, 1]$;
- *Expected distance* (Ed_o) is the expected value of the distance to the best unit attained by DMU_o : $Ed_o = \sum_{t=1}^T d_o^{(t)} / T$, where $d_o^{(t)}$ is the distance to the best DMU attained by DMU_o with sample t .

Example. Table 3 provides the distribution of efficiency distances to the best unit and the respective expected values for the twelve hospitals. For clarity of presentation, we used five buckets with equal widths. For four hospitals (H_2, H_5, H_9 , and H_{12}), the distance to the best unit for each sample is not greater than 0.2 ($DAII(DMU_o, [0.0, 0.2]) = 1$). This means that for the vast majority of feasible scenarios, their margin to the most efficient hospital is minor. When comparing these results with the extreme distances, for each of these hospitals, a distance greater than 0.2 is possible for some scenarios (e.g., $d_{H_2}^* = 0.343$). However, the distribution analysis shows that such cases are sporadic. Moreover, $DAIIs$ indicate that H_4 is the only hospital for which the distance to the best unit is likely to be worse than 0.4. This holds for more than half of the samples (58%). All other hospitals attain distances between 0.0 and 0.4 for all samples. The expected distances (Ed) can be used to order the hospitals from the best to the worst. In particular, H_5 is ranked at the top with $Ed_5 = 0.021$. It is followed by H_{12} , H_2 , and H_{11} , whereas H_4 is ranked at the bottom ($Ed_4 = 0.404$).

3.2. Scores

This section considers the comprehensive value scores E_o for DMU_o . They might be viewed as absolute efficiencies, independent of the other DMUs. To find the maximal and minimal scores for DMU_o , one

Table 3: Distribution and expected values of the distance to the best unit for the twelve hospitals in the illustrative study.

DMU	$[0.0 - 0.2]$	$(0.2 - 0.4]$	$(0.4 - 0.6]$	$(0.6 - 0.8]$	$(0.8 - 1]$	Ed_o
H_1	0.28	0.72	0.00	0.00	0.00	0.229
H_2	1.00	0.00	0.00	0.00	0.00	0.045
H_3	0.50	0.50	0.00	0.00	0.00	0.206
H_4	0.00	0.42	0.58	0.00	0.00	0.404
H_5	1.00	0.00	0.00	0.00	0.00	0.021
H_6	0.43	0.57	0.00	0.00	0.00	0.214
H_7	0.03	0.97	0.00	0.00	0.00	0.264
H_8	0.29	0.71	0.00	0.00	0.00	0.223
H_9	1.00	0.00	0.00	0.00	0.00	0.081
H_{10}	0.95	0.05	0.00	0.00	0.00	0.128
H_{11}	0.99	0.01	0.00	0.00	0.00	0.046
H_{12}	1.00	0.00	0.00	0.00	0.00	0.035

needs to solve the following models:

$$\text{Maximize } \sum_{q=1}^Q U_{q,o}, \text{ s.t. } \mathcal{W} \cup \mathcal{ORD}(OI, OO) \cup \mathcal{IMP}^{OPT}(II, IO); \quad (12)$$

$$\text{Minimize } \sum_{q=1}^Q U_{q,o}, \text{ s.t. } \mathcal{W} \cup \mathcal{ORD}(OI, OO) \cup \mathcal{IMP}^{PES}(II, IO). \quad (13)$$

Example. The extreme scores attained by the twelve hospitals in the illustrative study are shown in Table 4. The best maximal score (E_o^*) is attained by H_{12} (0.828), followed by H_{11} (0.815) and H_5 (0.808). On the contrary, the worst maximal score is observed for H_4 (0.572). The minimal score is the highest for H_{11} (0.360) and H_{12} (0.318). In turn, H_4 attains the worst minimal score (0.087).

Table 4: Extreme scores for the twelve hospitals in the illustrative study.

	H_1	H_2	H_3	H_4	H_5	H_6	H_7	H_8	H_9	H_{10}	H_{11}	H_{12}
$E_{o,*}$	0.203	0.250	0.178	0.087	0.230	0.186	0.173	0.183	0.259	0.250	0.360	0.318
E_o^*	0.707	0.792	0.735	0.572	0.808	0.735	0.701	0.716	0.784	0.771	0.815	0.828

When summarizing the outcomes of the stochastic analysis, we consider the following results:

- *(Absolute) Efficiency Acceptability Interval Index* ($EAI(DMU_o, b_i)$) is the share of feasible scenarios for which the score of DMU_o is in the interval $b_i \subseteq [0, 1]$;
- *Expected (absolute) efficiency* (EE_o) is the expected value of score for DMU_o , $EE_o = \sum_{t=1}^T E_o^{(t)} / T$.

Example. The distribution of scores given five buckets with equal widths is provided in Table 5. Hospital H_4 is the only one for which the score falls below 0.2 ($EAI(H_4, [0.0, 0.2]) = 0.41$). For all other hospitals, the scores for all samples are in the three intermediate intervals (between 0.2 and 0.8). We can discriminate the hospitals according to their score distribution. For example, the score of H_5 is better than 0.6 for most samples (52%), and it falls for no sample below 0.4. Hospital H_7 attains a score in the quite poor range (0.2, 0.4] for 81% scenarios. The expected scores (EEs) yield the same ranking of hospitals as the expected distances to the best unit (Eds). In particular, H_5 is ranked at the top with $EE_5 = 0.606$, and H_4 is ranked at the bottom ($EE_4 = 0.223$). The two expectation-based rankings are indeed equivalent: on any

simulated instance t , if some DMU_i has a score above DMU_j by a difference of $E_i^{(t)} - E_j^{(t)} = \delta$, then the distance between DMU_i and the best DMU under instance t will be less by a difference of δ than the distance between DMU_j and the same best DMU.

Table 5: Distribution of efficiency scores and expected efficiencies for the twelve hospitals in the illustrative study.

DMU	[0.0 – 0.2]	(0.2 – 0.4]	(0.4 – 0.6]	(0.6 – 0.8]	(0.8 – 1]	EE_o
H_1	0.00	0.53	0.47	0.00	0.00	0.397
H_2	0.00	0.00	0.65	0.35	0.00	0.582
H_3	0.00	0.33	0.66	0.01	0.00	0.421
H_4	0.41	0.59	0.00	0.00	0.00	0.223
H_5	0.00	0.00	0.48	0.52	0.00	0.606
H_6	0.00	0.43	0.57	0.00	0.00	0.412
H_7	0.00	0.81	0.19	0.00	0.00	0.363
H_8	0.00	0.50	0.50	0.00	0.00	0.404
H_9	0.00	0.00	0.90	0.10	0.00	0.545
H_{10}	0.00	0.01	0.98	0.01	0.00	0.499
H_{11}	0.00	0.00	0.65	0.35	0.00	0.581
H_{12}	0.00	0.00	0.60	0.40	0.00	0.591

3.3. Efficiency Ranks

The efficiency ranks derive from the ordinal comparison between DMUs, hence being more stable than the cardinal measures of efficiency when the DMU data are modified. To identify the best (minimal) rank for DMU_o , we minimize the number of other DMUs with score greater than the score of DMU_o in its most optimistic scenario:

$$\begin{aligned}
 & \text{Minimize } 1 + \sum_{k=1, \dots, K; k \neq o} b_k \\
 \text{s.t. } & \left. \begin{aligned}
 & \sum_{q=1}^Q U_{q,k} - \sum_{q=1}^Q U_{q,o} \leq C b_k, \text{ for } k = 1, \dots, K, k \neq o, \\
 & b_k \in \{0, 1\}, \text{ for } k = 1, 2, \dots, K, k \neq o, \\
 & \mathcal{W}, \text{ ORD}(OI, OO), \text{ IMP}^{OPT}(II, IO).
 \end{aligned} \right\} \quad (14)
 \end{aligned}$$

Each binary variable b_k attains a value of zero if $E_k \leq E_o$. Otherwise, the value of b_k is equal to 1, meaning that DMU_k is ranked better than DMU_o . Hence the sum of b_k , $k = 1, \dots, K$, $k \neq o$, increased by one corresponds to the rank of DMU_o .

To find the worst (maximal) rank of DMU_o , we maximize the number of other DMUs with scores not worse than the score of DMU_o :

$$\begin{aligned}
 & \text{Maximize } 1 + \sum_{k=1, \dots, K; k \neq o} b_k \\
 \text{s.t. } & \left. \begin{aligned}
 & \sum_{q=1}^Q U_{q,o} - \sum_{q=1}^Q U_{q,k} \leq C(1 - b_k), \text{ for } k = 1, \dots, K, k \neq o, \\
 & b_k \in \{0, 1\}, \text{ for } k = 1, 2, \dots, K, k \neq o, \\
 & \mathcal{W}, \text{ ORD}(OI, OO), \text{ IMP}^{PES}(II, IO).
 \end{aligned} \right\} \quad (15)
 \end{aligned}$$

Example. Table 6 provides the extreme ranks for the twelve hospitals. The five previously deemed efficient hospitals attain the best possible rank of 1. Among the remaining units, H_{10} is the best with $R_{o^*} = 2$ and H_4 is the worst with $R_{o^*} = 11$. Regarding the worst possible rank (R_o^*), all hospitals are ranked outside the

top seven for at least one scenario. The best among them is H_2 with $R_o^* = 8$, and three other hospitals are not ranked outside the top ten ($R_{12}^* = 9$, $R_5^* = R_{11}^* = 10$). Most DMUs attain at most 11-th rank. There are only two hospitals (H_4 and H_6) which, in their pessimistic scenario, are ranked at the very bottom.

Table 6: Extreme efficiency ranks for the twelve hospitals in the illustrative study.

	H_1	H_2	H_3	H_4	H_5	H_6	H_7	H_8	H_9	H_{10}	H_{11}	H_{12}
R_{o^*}	3	1	3	11	1	3	3	3	1	2	1	1
R_o^*	11	8	11	12	10	12	11	11	11	11	10	9

The most interesting rank-related stochastic outcomes are defined in the following way:

- *Efficiency Rank Acceptability Index* ($ERAI(DMU_o, r)$) is the share of feasible scenarios for which DMU_o attains r -th position in the ranking of DMUs;
- *Expected rank* (ER_o) is the estimated expected rank for DMU_o computed as follows: $ER_o = \sum_{k=1}^K k \cdot ERAI(DMU_o, k)$.

Example. Table 7 presents the $ERAI$ s for the twelve hospitals, providing information on the distribution of efficiency ranks for all feasible scenarios. For example, the possible ranks obtained for H_1 with mathematical programming are in the range [3, 11], while the stochastic analysis indicates the positive probabilities only for ranks between 7 and 11. This means that H_1 can attain ranks between 3 and 6, but it is improbable. The most common rank for this hospital is 10-th ($ERAI(H_1, 10) = 0.4$). Among the five efficient hospitals, H_5 is ranked at the top for most samples (59%), and H_9 is not ranked first for any sample. $ERAI$ s indicate the subsets of hospitals for which the efficiency ranks are rather stable and those for which the attained positions strongly depend on the selected weight vector, precise performances, and marginal value functions. For example, H_4 is ranked last for all samples, which confirms its poor performance. Similarly, H_{10} attains ranks between 5 and 6, while being placed sixth for 88% samples. On the contrary, the ranks of H_{11} are distributed between 1 and 6, with no $ERAI$ greater than 26%, suggesting that the position of this DMU strongly depends on the considered instance. The analysis of expected ranks lets us determine a complete ranking of hospitals, with H_5 (1.79), H_{12} (2.49), and H_{11} (2.97) at the podium, and H_4 (12.00) and H_7 (10.91) at the bottom.

Table 7: Efficiency rank acceptability indices and expected ranks for the twelve hospitals in the illustrative study.

DMU	Rank												ER_o
	1	2	3	4	5	6	7	8	9	10	11	12	
H_1	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.25	0.24	0.40	0.02	0.00	9.01
H_2	0.01	0.36	0.18	0.42	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	3.11
H_3	0.00	0.00	0.00	0.00	0.00	0.00	0.46	0.41	0.12	0.01	0.00	0.00	7.68
H_4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	12.00
H_5	0.59	0.12	0.22	0.05	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.79
H_6	0.00	0.00	0.00	0.00	0.00	0.00	0.41	0.12	0.14	0.29	0.04	0.00	8.43
H_7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.07	0.92	0.00	10.91
H_8	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.22	0.49	0.23	0.02	0.00	8.97
H_9	0.00	0.00	0.08	0.15	0.70	0.07	0.00	0.00	0.00	0.00	0.00	0.00	4.76
H_{10}	0.00	0.00	0.00	0.00	0.12	0.88	0.00	0.00	0.00	0.00	0.00	0.00	5.88
H_{11}	0.26	0.15	0.17	0.24	0.14	0.04	0.00	0.00	0.00	0.00	0.00	0.00	2.97
H_{12}	0.14	0.37	0.35	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.49

3.4. Preference Relations

When comparing the DMUs pairwise, it is possible to verify two types of preference relations. The possible one \succsim_E^P holds for a pair (DMU_o, DMU_k) if $E_o \geq E_k$ for at least one feasible scenario. To verify its truth, we maximize the difference between the efficiencies E_o and E_k under the optimistic scenario for DMU_o :

Maximize $d_{o,k}$

$$\text{s.t.} \quad \left. \begin{array}{l} \sum_{q=1}^Q U_{q,o} - \sum_{q=1}^Q U_{q,k} \geq d_{o,k}, \\ \mathcal{W}, \text{ORD}(OI, OO), \text{IMP}^{OPT}(II, IO). \end{array} \right\} \quad (16)$$

If such a difference is greater than or equal to zero, $DMU_o \succsim_E^P DMU_k$. One can note that if this difference is negative for some DMU_k , i.e., if $DMU_o \succsim_E^P DMU_k$ does not hold, then DMU_o cannot be efficient and the optimal value of problem (10) will be greater than zero.

In turn, the necessary relation \succsim_E^N holds for a pair (DMU_o, DMU_k) if DMU_o attains efficiency at least as good as E_k for all feasible scenarios. To verify its truth, we minimize the difference between the efficiencies E_o and E_k under the pessimistic scenario for DMU_o :

Minimize $d_{o,k}$

$$\text{s.t.} \quad \left. \begin{array}{l} \sum_{q=1}^Q U_{q,o} - \sum_{q=1}^Q U_{q,k} \leq d_{o,k}, \\ \mathcal{W}, \text{ORD}(OI, OO), \text{IMP}^{PES}(II, IO). \end{array} \right\} \quad (17)$$

If such a minimal difference is not lesser than zero, $DMU_o \succsim_E^N DMU_k$.

Example. The necessary and possible preference relations for all pairs of hospitals are presented in Table 8. To visualize the necessary preference relation, we depict the respective Hasse diagram in Figure 2. Regarding the most robust preferences, H_2 and H_{12} are necessarily preferred to the highest number of other hospitals (4 and 3, respectively), while H_4 and H_6 are not necessarily preferred to any other hospital. Hospital H_4 is worse than 10 other units independently of the chosen weights, performances, and value functions.

Table 8: The necessary (N) and possible (P) preference relation for all pairs of hospitals in the illustrative study.

	H_1	H_2	H_3	H_4	H_5	H_6	H_7	H_8	H_9	H_{10}	H_{11}	H_{12}
H_1	N		P	N	P	P	P	P	P	P	P	P
H_2	N	N	N	N	P	P	P	N	P	P	P	P
H_3	P		N	N	P	P	P	P	P	P	P	P
H_4				N		P						
H_5	P	P	P	N	N	P	P	N	P	P	P	P
H_6	P	P	P	P	P	N	P	P	P	P		
H_7	P	P	P	N	P	P	N	P	P	P	P	P
H_8	P		P	N		P	P	N	P	P	P	P
H_9	P	P	P	N	P	P	P	P	N	P	P	P
H_{10}	P	P	P	N	P	P	P	P	P	N	P	
H_{11}	P	P	P	N	P	N	P	P	P	P	N	P
H_{12}	P	P	P	N	P	N	P	P	P	N	P	N

The major pair-oriented outcome derived from the stochastic analysis is the *Pairwise Efficiency Out-ranking Index* ($PEOI(DMU_o, DMU_k)$). It is defined as the share of feasible scenarios for which the

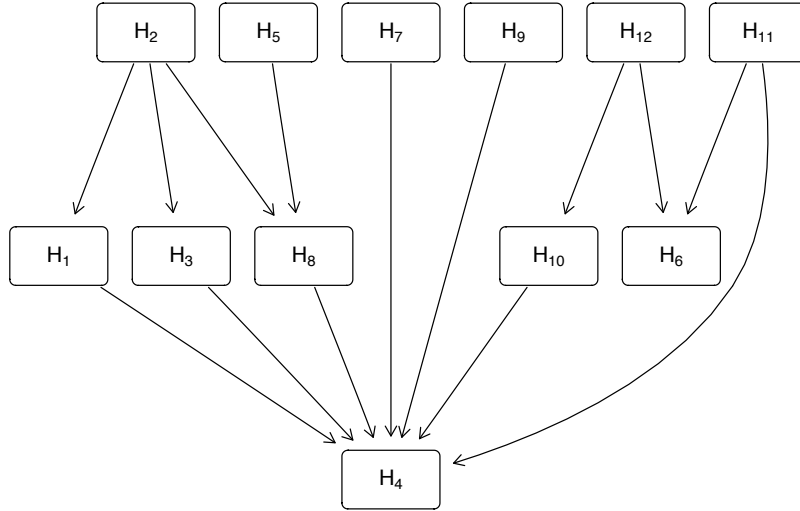


Figure 2: The Hasse diagram of the necessary efficiency preference relations for the illustrative study.

efficiency of DMU_o is not worse than the efficiency of DMU_k .

Example. The analysis of *PEOIs* (see Table 9) is particularly helpful for pairs not related by the necessary preference. Among them are pairs for which preference in one direction occurs more often than in the other direction. In particular, $PEOI(H_{10}, H_3) = 0.99$ confirms that H_{10} attains better efficiency than H_3 for the vast majority of scenarios. Similar observation holds for pairs (H_8, H_7) and (H_9, H_{10}) . For other pairs of hospitals, *PEOIs* are more balanced, making it challenging to indicate a better performer (see, e.g., (H_{12}, H_5) with $PEOI(H_{12}, H_5) = 0.41$ and (H_{11}, H_2) with $PEOI(H_{11}, H_2) = 0.43$). Obviously, for pairs with one hospital being necessarily preferred to the other, *PEOIs* are equal to one.

Table 9: Pairwise efficiency outranking indices for all pairs of hospitals in the illustrative study.

<i>DMU</i>	H_1	H_2	H_3	H_4	H_5	H_6	H_7	H_8	H_9	H_{10}	H_{11}	H_{12}
H_1	1.00	0.00	0.18	1.00	0.00	0.37	0.99	0.51	0.00	0.00	0.00	0.00
H_2	1.00	1.00	1.00	1.00	0.03	1.00	1.00	1.00	0.92	0.98	0.57	0.40
H_3	0.82	0.00	1.00	1.00	0.00	0.70	0.98	0.90	0.00	0.01	0.00	0.00
H_4	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
H_5	1.00	0.97	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00	0.67	0.59
H_6	0.63	0.00	0.30	1.00	0.00	1.00	0.94	0.58	0.00	0.00	0.00	0.00
H_7	0.01	0.00	0.02	1.00	0.00	0.06	1.00	0.02	0.00	0.00	0.00	0.00
H_8	0.49	0.00	0.10	1.00	0.00	0.42	0.98	1.00	0.00	0.00	0.00	0.00
H_9	1.00	0.08	1.00	1.00	0.02	1.00	1.00	1.00	1.00	0.95	0.25	0.04
H_{10}	1.00	0.02	0.99	1.00	0.00	1.00	1.00	1.00	0.05	1.00	0.03	0.00
H_{11}	1.00	0.43	1.00	1.00	0.33	1.00	1.00	1.00	0.75	0.97	1.00	0.31
H_{12}	1.00	0.60	1.00	1.00	0.41	1.00	1.00	1.00	0.96	1.00	0.69	1.00

4. Software Implementation

The methods for robustness analysis in value-based additive efficiency analysis have been implemented in R. They have been originally designed for use on the *diviz* platform [20]. However, they can also be employed as independent programming modules. The source code of these approaches is available at https://github.com/alabijak/diviz_DEA/tree/master/ImpreciseDEAValueADD. It is divided into the following components:

- *ImpreciseDEA-ValueAdditive_efficiencies* produces extreme efficiency scores and distances to the best unit for all DMUs;
- *ImpreciseDEA-ValueAdditive_extremeRanks* provides the extreme ranks for each DMU;
- *ImpreciseDEA-ValueAdditive_preferenceRelations* identifies all pairs of DMUs for which the necessary or possible efficiency preference relation holds;
- *ImpreciseDEA-ValueAdditive-SMAA_efficiencies* implements the Monte Carlo simulation to determine the efficiency distribution, the expected efficiency scores (EE_o), and the extreme efficiencies attained in the analyzed sample;
- *ImpreciseDEA-ValueAdditive-SMAA_ranks* computed the efficiency rank acceptability indices and the expected rank for each unit;
- *ImpreciseDEA-ValueAdditive-SMAA_preferenceRelations* calculates the pairwise efficiency outranking indices for all pairs of DMUs.

The modules mentioned above accept the following input files describing the analyzed problem:

- *units*: the list of DMUs;
- *inputs/outputs*: the description of inputs and outputs, containing additional information about the scales of all factors (quantitative or ordinal) and, optionally, the specification of per-factor marginal value functions; if a pair of functions are provided, they are considered to be the lower and upper bounds for the admissible function range, if only one value function is given, it is treated as a precise value function for a given factor; if no function is provided, we assume it to be linear;
- *performance*: the lower bounds of the performance intervals assigned to all DMUs and all factors (they are interpreted as precise performances if the *max performance* file is not provided);
- *max performance* (optional): the upper bounds of the performance intervals assigned to all units and all factors;
- *weight constraints* (optional): the absolute or relative linear constraints defining the space of feasible weights.

The modules also process some additional parameters:

- *tolerance* (by default: 0) allows to construct the interval performances from the precise ones using the following transformation: $[(1 - tolerance) \cdot x_{q,k}; (1 + tolerance) \cdot x_{q,k}]$ for inputs and $[(1 - tolerance) \cdot y_{q,k}; (1 + tolerance) \cdot y_{q,k}]$ for outputs;
- *transform to utilities* (by default: “Yes”) is the Boolean parameter indicating if the provided performances should be transformed using the marginal value functions; if this parameter is set to “No”, then the input performances should already be given in the form of marginal values;
- *function shapes provided* (by default: “No”) provides information if the shapes of marginal value functions are provided in the *inputs/outputs* file; if not, then linear marginal value functions are used in calculations;

- *boundaries provided* (default: “No”) is the parameter used only if the shapes of marginal value functions are not provided; it refers to the specification of the extreme performances; if the parameter is set to “Yes”, then the boundaries are the lower and upper limits of the linear value functions; if it is set to “No”, the boundaries are automatically derived from the minimal and maximal performances of DMUs observed for a given factor;
- *number of samples* (only for simulation-based modules; by default: 100) specifies the number of samples used in the Monte Carlo simulation;
- *number of intervals* (only for *ImpreciseDEA-ValueAdditive-SMAA-efficiencies*; by default: 10) is the number of intervals (buckets) used in the calculation of distance- and score-based acceptability indices.

5. Case Study

In this section, we present the results of applying the proposed framework to the real-world case study concerning the evaluation of Special Economic Zones (SEZ) in Poland. SEZ is a dedicated area of the country’s territory where the business can be run under preferential conditions. The main reasons for creating such zones are accelerating regions’ development, managing post-industrial property and infrastructure, creating new jobs, and attracting foreign investors. In Polish SEZs, the assistance is allocated in the form of income tax and real property tax exemptions. Specifically, the company does not pay income taxes for the earnings between the permit’s date and the exhaustion of regional aid or the end of an SEZ’s operation. Regional aid may also be allocated in the form of a real property tax exemption introduced by a municipal board, which needs to adopt a relevant resolution. The special economic zones in Poland were established in 1994, and over a few decades of their existence, they have grown to 20,000 hectares of area with investments at the level of over 100 billion PLN a year and creating over 200 000 new jobs.

We analyze the performance of fourteen Special Economic Zones listed in Table 10. The relevant data is computed based on the indicators reported by the Polish Ministry of Entrepreneurship and Technology [24]. In particular, we consider two inputs:

- i_1 , *area* (interval): a total area of the SEZ in years 2016–2017 (ha),
- i_2 , *expenditures*: the capital expenditures made by investors in a given SEZ at the end of 2017 (millions of PLN),

and two outputs:

- o_1 , *jobs* (interval): the number of jobs in years 2016–2017,
- o_2 , *financial_result*: the financial result of the management companies in 2017 (thousands of PLN).

In the data, the intervals for i_1 and o_1 reflect variations throughout 2016 and 2017. For each factor, we elicited extreme marginal value functions from the expert on the functioning of SEZs in Poland (see Figure 3). Moreover, to prevent any factor’s negligible and dominating role, we have restricted the weights of all input and output value functions to the range $[1/6, 1/3]$. In what follows, we discuss the results of the robustness analysis.

Table 10: The performances of fourteen Special Economic Zones on two inputs and two outputs.

SEZ	Short name	Area (i_1)	Expenditures (i_2)	Jobs (o_1)	Financial result (o_2)
Kamienna Góra	KAM	[373.83, 540.83]	2557.3	[7347, 7530]	555.1
Katowice	KAT	[2614.40, 2614.40]	16605.1	[59964, 64481]	17663.5
Kostrzyn-Słubice	KOS	[1936.90, 2201.25]	7133.4	[31927, 32400]	22984.9
Kraków	KRA	[866.80, 949.66]	4240.4	[25862, 29580]	1373.0
Legnica	LEG	[1341.15, 1341.15]	5131.8	[14367, 15294]	7614.5
Łódź	LOD	[1416.84, 1754.64]	13318.7	[33401, 36122]	7402.8
Mielec	MIE	[1643.12, 1723.97]	7838.1	[24815, 34992]	4956.0
Pomorze	POM	[2246.29, 2246.29]	10481.6	[22921, 24893]	1479.1
Słupsk	SLU	[910.16, 910.16]	1592.3	[3478, 3941]	761.5
Starachowice	STA	[664.16, 707.98]	1790.9	[6829, 7260]	701.0
Suwałki	SUW	[635.07, 662.95]	2500.1	[7258, 8336]	2734.4
Tarnobrzeg	TAR	[1868.21, 1868.21]	7470.7	[20740, 23734]	18220.4
Wałbrzych	WAL	[3554.96, 3774.55]	22789.5	[48954, 50268]	11862.8
Warmia-Mazury	WAR	[1364.68, 1390.73]	3124.6	[17643, 20778]	1647.4

5.1. Distances to the best unit and scores

This section presents the distance- and score-based results. In particular, the extreme values of these measures are presented in Table 11. Eight SEZs (KAM, KOS, KRA, LOD, MIE, SUW, TAR, and WAR) perform efficiently in their best scenario ($d_{*,o} = 0$). The remaining six SEZs attain the minimal distances to the best unit greater than zero. Among these inefficient zones, there are significant differences between the results they attain for the most favorable distance values. In particular, STA is very close to being efficient ($d_{*,STA} = 0.03$), whereas WAL is far from efficiency ($d_{*,WAL} = 0.311$). The minimum among the maximal distances to the best zone is attained by KOS ($D_{KOS}^* = 0.064$). Hence, this SEZ is nearly efficient even in its least favorable scenario and it can be deemed as the overall best performer. The second best zone, according to the maximal distance, is TAR ($D_{TAR}^* = 0.374$), whereas the worst maximal distances are attained by SLU (0.609) and WAL (0.733).

When it comes to the value scores (absolute efficiencies), the zones with the greatest optimistic scores (E_o^*) are KOS (0.916), KRA (0.765), and TAR (0.740). The worst maximal score is observed for WAL. This SEZ's efficiency does not exceed 0.383 even in the most favorable scenario. The minimal possible score is also the most favorable for KOS (0.678), which confirms that this zone performs well for all feasible scenarios. Other zones with minimal scores greater than 0.5 are KRA (0.508) and TAR (0.507).

Table 11: Extreme and expected values of distances to the efficient unit and scores for Special Economic Zones.

	$d_{*,o}$	d_o^*	Ed_o	E_o^*	$E_{*,o}$	EE_o
KAM	0.000	0.553	0.264	0.686	0.363	0.520
KAT	0.060	0.546	0.316	0.637	0.349	0.467
KOS	0.000	0.064	0.000	0.916	0.678	0.784
KRA	0.000	0.408	0.158	0.765	0.508	0.626
LEG	0.020	0.524	0.254	0.670	0.384	0.530
LOD	0.000	0.501	0.284	0.712	0.343	0.500
MIE	0.000	0.465	0.246	0.739	0.420	0.537
POM	0.165	0.548	0.343	0.607	0.340	0.440
SLU	0.039	0.609	0.316	0.639	0.307	0.468
STA	0.003	0.560	0.273	0.675	0.356	0.511
SUW	0.000	0.560	0.264	0.680	0.356	0.520
TAR	0.000	0.374	0.200	0.740	0.507	0.583
WAL	0.311	0.733	0.522	0.383	0.150	0.262
WAR	0.000	0.472	0.209	0.715	0.441	0.575

The analysis of extreme values can be enriched with the stochastic distribution of the respective re-

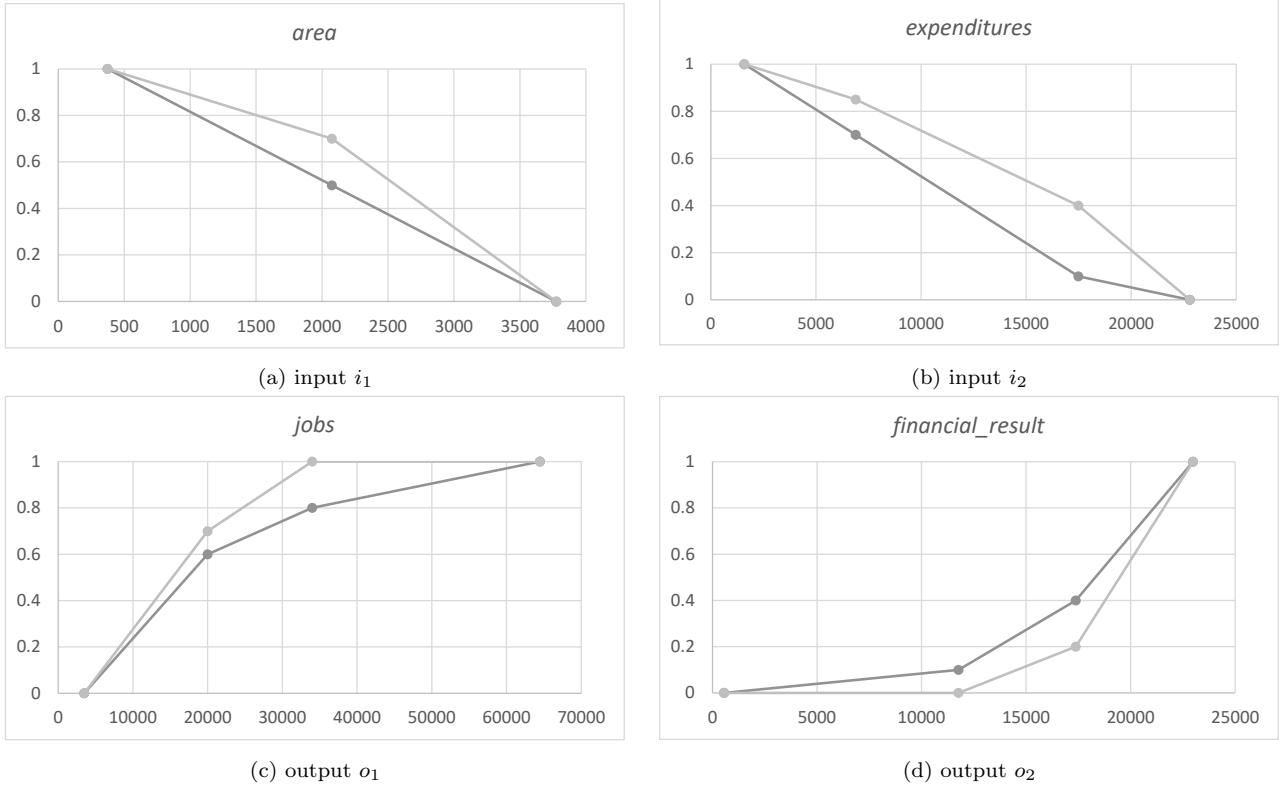


Figure 3: Admissible function ranges for inputs and outputs in the analyzed data set.

sults. In Table 12, we provide the distance-oriented acceptability indices given ten intervals. They are estimated based on 10,000 samples. For KOS, the distance to the best zone is never greater than 0.1 ($DAII(KOS, [0.0 - 0.1]) = 1$), confirming that it is robustly close to being efficient. Similarly, for KRA, for more than 75% of weights, performances, and marginal value functions, the distance to the best SEZ does not exceed 0.2. On the contrary, the worst zone is WAL. For most samples (66%), its distance to the best zone is greater than 0.5. The stochastic analysis also gives the expected distances to the best SEZ (see Table 11). The best zones according to this measure (Ed_o) are KOS (0.000), KRA (0.158), and TAR (0.200), while the least favorable ones are WAL (0.522) and POM (0.343).

Due to the high correlation between distances and efficiencies, we omit the detailed results for the distribution of scores. Instead, let us note that KOS attains a score of at least 0.7 for all samples, WAL attains a score of at most 0.4 for all samples, and the best expected efficiencies are associated with KOS (0.784), KRA (0.626), and TAR (0.583).

5.2. Efficiency ranks

When it comes to the efficiency ranks, we report the extreme positions, the efficiency rank acceptability indices, and the expected ranks for all SEZs in Table 13. Naturally, the zones classified as efficient attain the first rank in their most favorable scenarios. Among the inefficient zones, the best ranks are attained by KAT and STA ($R_{*,KAT} = R_{*,STA} = 2$). The worst SEZ considering R_* is WAL, ranked at most ninth. When considering the worst ranks, only WAL and SLU fall into the last position ($R^* = 14$). The best pessimistic ranks are observed for KOS (6) and KRA (8). However, the efficient zones (LOD, MIE, and SUW) can be ranked in the second-last position in their least favorable scenario. This indicates that their ranking strongly depends on the chosen weight vector, precise performances, and marginal value functions.

Table 12: Distribution of the distances to the best unit for Special Economic Zones.

SEZ	[0.0, 0.1]	(0.1, 0.2]	(0.2, 0.3]	(0.3, 0.4]	(0.4, 0.5]	(0.5, 0.6]	(0.6, 0.7]	(0.7, 0.8]	(0.8, 0.9]	(0.9, 1]
KAM	0.003	0.152	0.555	0.277	0.013	0.000	0.000	0.000	0.000	0.000
KAT	0.000	0.002	0.355	0.619	0.024	0.000	0.000	0.000	0.000	0.000
KOS	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
KRA	0.172	0.587	0.241	0.000	0.000	0.000	0.000	0.000	0.000	0.000
LEG	0.000	0.150	0.641	0.208	0.001	0.000	0.000	0.000	0.000	0.000
LOD	0.000	0.046	0.560	0.391	0.003	0.000	0.000	0.000	0.000	0.000
MIE	0.000	0.196	0.658	0.146	0.000	0.000	0.000	0.000	0.000	0.000
POM	0.000	0.000	0.186	0.708	0.106	0.000	0.000	0.000	0.000	0.000
SLU	0.000	0.044	0.352	0.512	0.092	0.000	0.000	0.000	0.000	0.000
STA	0.003	0.128	0.527	0.323	0.019	0.000	0.000	0.000	0.000	0.000
SUW	0.003	0.152	0.559	0.273	0.013	0.000	0.000	0.000	0.000	0.000
TAR	0.002	0.487	0.511	0.000	0.000	0.000	0.000	0.000	0.000	0.000
WAL	0.000	0.000	0.000	0.002	0.338	0.623	0.037	0.000	0.000	0.000
WAR	0.009	0.430	0.519	0.042	0.000	0.000	0.000	0.000	0.000	0.000

Table 13: Extreme and expected ranks and efficiency rank acceptability indices for Special Economic Zones.

	$R_{s,o}$	R_o^*	ER_o	Rank													
				1	2	3	4	5	6	7	8	9	10	11	12	13	14
KAM	1	13	7.137	0.000	0.000	0.016	0.054	0.165	0.150	0.173	0.176	0.148	0.101	0.017	0.000	0.000	0.000
KAT	2	13	10.770	0.000	0.000	0.000	0.015	0.025	0.048	0.049	0.052	0.029	0.056	0.205	0.311	0.210	0.000
KOS	1	6	1.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
KRA	1	8	2.065	0.000	0.935	0.065	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
LEG	3	13	6.804	0.000	0.000	0.000	0.010	0.177	0.277	0.167	0.282	0.078	0.009	0.000	0.000	0.000	0.000
LOD	1	13	9.123	0.000	0.000	0.000	0.006	0.039	0.110	0.146	0.057	0.057	0.227	0.316	0.039	0.003	0.000
MIE	1	13	6.662	0.000	0.000	0.014	0.084	0.333	0.143	0.066	0.055	0.197	0.102	0.006	0.000	0.000	0.000
POM	3	13	12.379	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.019	0.013	0.055	0.396	0.517	0.000
SLU	6	14	11.315	0.000	0.000	0.000	0.000	0.000	0.000	0.003	0.008	0.078	0.238	0.198	0.205	0.270	0.000
STA	2	12	8.999	0.000	0.000	0.000	0.002	0.006	0.055	0.159	0.166	0.210	0.171	0.182	0.049	0.000	0.000
SUW	1	13	7.409	0.000	0.000	0.003	0.014	0.116	0.183	0.214	0.186	0.181	0.082	0.021	0.000	0.000	0.000
TAR	1	12	3.552	0.000	0.064	0.549	0.283	0.048	0.014	0.020	0.018	0.003	0.001	0.000	0.000	0.000	0.000
WAL	9	14	14.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000
WAR	1	12	3.785	0.000	0.001	0.353	0.532	0.091	0.020	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.000

The calculated $ERAI$ s allow observing how the ranks are distributed for the feasible scenarios. For example, even though the exact analysis indicates that KOS can be ranked between first and sixth, it is ranked at the top for each sample ($ERAI(KOS, 1) = 1$). Hence even if there are scenarios with five other SEZs better than KOS, such an outcome is highly improbable. On the contrary, WAL is ranked last for all samples. The ranks of other SEZs also exhibit high stability. For example, KRA is ranked second for 93.5% samples, while POM attains positions between 12 and 13 for 91.3% scenarios. However, there are also some zones for which the ranks are more distributed. In particular, SUW and KAM attain ranks between 3 and 11, with no $ERAI$ exceeding 21.4% for SUW and 17.3% for KAM. The stochastic analysis can also be used to rank all SEZs in line with their expected ranks. In this case, KOS (1), KRA (2.065), and TAR (3.552) prove to be the best, whereas SLU (11.315), POM (12.379), and WAL (14.000) are ranked at the bottom.

5.3. Preference relations

As far as the preference relations are concerned, Table 14 indicates the pairs of SEZs for which the possible (P) and necessary (N) relations hold. Also, the necessary preference relation is presented graphically in Figure 4. Obviously, the eight efficient zones are possibly preferred to all remaining ones. Among the inefficient ones, KAT and STA are possibly preferred to twelve other SEZs. In turn, WAL performs the least favorably, being possibly preferred only to five other zones. As for the necessary preference relation, KOS is robustly preferred to the greatest number (6) of other zones, while POM, SLU, and WAL are not

necessarily preferred to any other SEZ. The zones which are necessarily worse than the greatest number of other SEZs are WAL (8) and SLU (5).

Table 14: The necessary (N) and possible (P) efficiency preference relations for all pairs of Special Economic Zones.

	KAM	KAT	KOS	KRA	LEG	LOD	MIE	POM	SLU	STA	SUW	TAR	WAL	WAR
KAM	N	P	P	P	P	P	P	P	N	P	P	P	P	P
KAT	P	N		P	P	P	P	P	P	P	P	P	N	P
KOS	P	N	N	P	N	P	P	N	N	N	P	P	N	P
KRA	P	P	P	N	N	P	P	N	N	P	P	P	N	P
LEG	P	P			N	P	P	P	P	P	P	P	N	P
LOD	P	P	P	P	P	N	P	P	P	P	P	P	N	P
MIE	P	P	P	P	P	P	N	P	P	P	P	P	N	P
POM	P	P			P	P	P	N	P	P	P	P	P	P
SLU		P			P	P	P	P	N			P	P	P
STA	P	P		P	P	P	P	P	N	N	P	P	P	P
SUW	P	P	P	P	P	P	P	P	N	P	N	P	P	P
TAR	P	P	P	P	P	P	P	P	P	P	P	N	N	P
WAL	P							P	P	P	P		N	
WAR	P	P	P	P	P	P	P	P	P	P	P	P	N	N

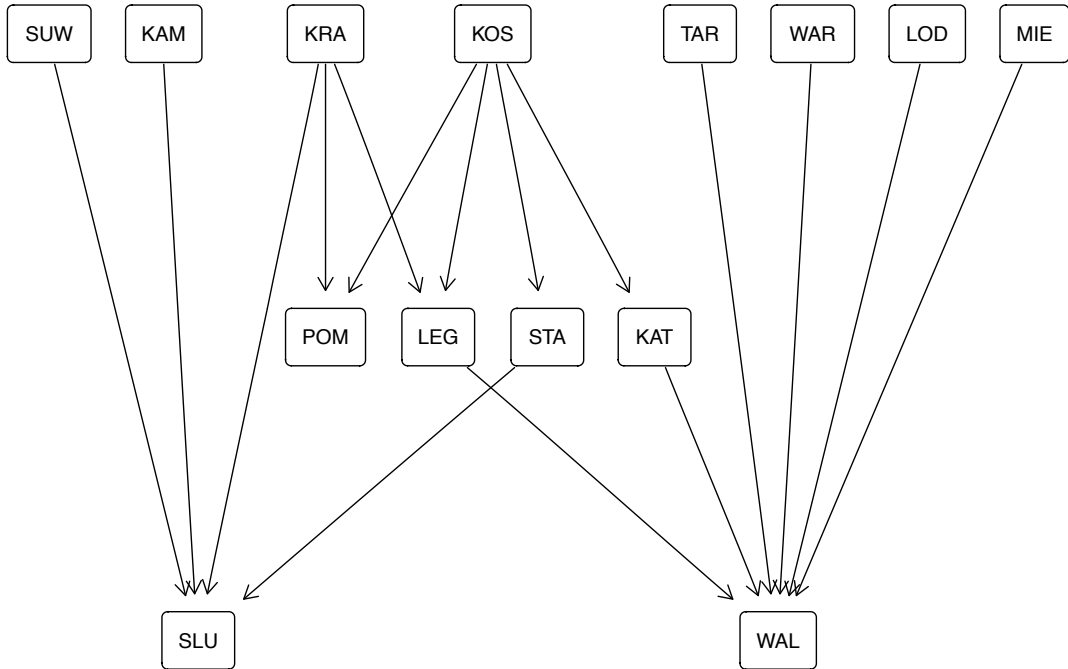


Figure 4: The Hasse diagram of the necessary efficiency preference relation for Special Economic Zones.

The Pairwise Efficiency Outranking Indices (see Table 15) provide detailed insight into the performance comparison for pairs not related by the necessary preference. For some pairs, one SEZ is better than the other for the vast majority of scenarios. They include, e.g., (WAR, KAT) and (KRA, MIE), for which the advantage of the first zone in the ordered pair is very high. In the same spirit, the *PEOIs* for pairs involving KOS and all other zones are equal to one. Thus, even if this zone is not necessarily preferred to all other SEZs, it proved to be at least as good for all samples. However, there are also some pairs of SEZs with more balanced performances. For example, the stochastic acceptabilities for (KAM, SUW), (STA, LOD), and (SLU, KAT) are not greater than 60% in any direction.

Table 15: Paiwise efficiency outranking indices for all pairs of Special Economic Zones.

	KAM	KAT	KOS	KRA	LEG	LOD	MIE	POM	SLU	STA	SUW	TAR	WAL	WAR
KAM	1.000	0.767	0.000	0.000	0.371	0.647	0.335	0.965	1.000	0.943	0.552	0.062	1.000	0.043
KAT	0.233	1.000	0.000	0.000	0.105	0.157	0.017	0.781	0.485	0.270	0.231	0.000	1.000	0.016
KOS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
KRA	1.000	1.000	0.000	1.000	1.000	1.000	0.999	1.000	1.000	1.000	1.000	0.915	1.000	0.999
LEG	0.629	0.895	0.000	0.000	1.000	0.773	0.325	1.000	0.995	0.762	0.649	0.025	1.000	0.005
LOD	0.353	0.843	0.000	0.000	0.227	1.000	0.052	0.978	0.663	0.410	0.360	0.004	1.000	0.013
MIE	0.665	0.983	0.000	0.001	0.675	0.948	1.000	1.000	0.947	0.738	0.675	0.039	1.000	0.117
POM	0.035	0.219	0.000	0.000	0.000	0.022	0.000	1.000	0.306	0.055	0.032	0.000	1.000	0.000
SLU	0.000	0.515	0.000	0.000	0.005	0.337	0.053	0.694	1.000	0.000	0.000	0.003	0.998	0.000
STA	0.057	0.730	0.000	0.000	0.238	0.590	0.262	0.945	1.000	1.000	0.066	0.035	1.000	0.007
SUW	0.448	0.769	0.000	0.000	0.351	0.640	0.325	0.968	1.000	0.934	1.000	0.049	1.000	0.022
TAR	0.938	1.000	0.000	0.085	0.975	0.996	0.961	1.000	0.997	0.965	0.951	1.000	1.000	0.661
WAL	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.000	0.000	0.000	1.000	0.000
WAR	0.957	0.984	0.000	0.001	0.995	0.987	0.883	1.000	1.000	0.993	0.978	0.339	1.000	1.000

5.4. Practical conclusions derived from the outcomes of robustness analysis

Administrative authorities may use the results of robustness analysis to rank the economic zones, compare their performances, or find realistic scenarios to improve the inefficient ones. In what follows, we discuss the major practical conclusions that can be derived from various outcomes.

We can discriminate between the efficient SEZs, attaining $d_{*,o} = 0$, based on their worst possible distances (d_o^*) to the best zone. For our problem, there are eight efficient zones. However, three (KAM, LOD, and SUW) attain distances to the best SEZ greater than 0.5 in their least favorable scenarios. Hence they can be judged worse than other efficient SEZs. For example, the pessimistic distance of KOS is vastly lesser (0.064). Such analysis motivates a detailed consideration of efficient zones' input and output performances, indicating the ones requiring improvements. For example, the inputs and outputs of SUW are low (e.g., only SLU has lower expenditures, while the area is lesser only for KAM). Consequently, when the weights assigned to outputs are high compared to those associated with inputs, the efficiency of SUW becomes far from the best zone. Therefore, aiming to decrease the maximal distance of SUW to the best zone, one should focus on improving its financial results and/or jobs without significantly increasing its resources.

Furthermore, the stochastic results are perfect for indicating the overall good or poor performers. For example, KOS can be deemed superior to other SEZs as its estimated expected distance to the best zone is zero. Hence it attains the best results for all sampled scenarios even if the exact outcomes derived from optimization show that an extreme scenario exists where seven other zones are better. Such a highly favorable performance of KOS is implied by its significantly high outputs and relatively low inputs. In particular, its financial result is almost 23 million PLN, compared to 18.2 million for the second SEZ (WAL) and 7.1 million on average. Moreover, its expenditures (7133.4) are lower than the average (7612.5) observed in the set of all zones. Analyzing the expected distances or efficiencies allows for determining the good performers among the inefficient SEZs. In the average case, LEG, judged inefficient, proves better than three units deemed efficient (KAM, SUW, LOD). This suggests that attaining average performances on all inputs and outputs is better than performing exceptionally well on one factor and poorly on the remaining ones.

Similarly, the robustness analysis hints at the stability of SEZs' performance given the cardinal and ordinal measures, i.e., efficiencies and ranks. In particular, KOS is robustly ranked first, and KRA is ranked second for 93.5% scenarios. In the case of KRA, this can be explained by its low inputs (only

949.66 ha of area at maximum and 4240.4 million of PLN expenditures) and one output which is relatively good when compared to other SEZs (i.e., at least 25862 jobs compared to 24468 on average). Conversely, WAL is ranked at the bottom for all samples. Its very high inputs (about twice the average for the area, and its expenditures are almost three times greater than the average) cannot compensate for outputs being only slightly higher than the average. The performance of other SEZs is less stable. For example, KAM is ranked in the range [1, 13], and its *ERAI*s are positive for all positions between 3 and 11. One of the underlying reasons is its relatively wide area interval, varying between 373.83 and 540.83.

The pairwise outcomes are beneficial if the analysts know some zones well and want to compare them with others. For example, if they are familiar with STA, they can note that SLU is always worse and KOS is robustly better. Also, the necessary preference graph allows for determining the possible improvement paths for inefficient zones. For example, WAL can choose among multiple such paths (see Figure 4). Some of them lead directly to TAR or WAR. Others suggest achieving first the efficiency of LEG. Hence the manager of WAL should notice that all inputs and outputs of WAL are greater than for LEG, but the difference in inputs is much greater than for outputs. In particular, the expenditures of WAL are much greater than those of LEG, while its financial result is only slightly better. Consequently, one should aim at decreasing the expenditures of WAL or significantly increasing its financial results. Having attained the level of LEG, one should then design ways to reach performances of KOS or KRA.

6. Summary

This paper introduced a novel framework for efficiency analysis, exploiting the multiplicity of relevant scenarios. The assumed model is an additive value function, aggregating input and output performances. The framework admits four types of imprecision concerning performances and preferences. In particular, it accepts interval assessments, ordinal judgments, linear constraints on the factors' weights, and specification of acceptable, marginal value functions via the range delimited by the pre-defined extreme shapes.

The proposed computational procedures combine mathematical programming with Monte Carlo simulations to derive various results. They concern four perspectives: the distance to an efficient unit, efficiency scores, ranks, and pairwise relations. For each of them, we capture different certainty levels referring to the necessary, possible, extreme, and expected viewpoints, as well as the distribution of outcomes in the space of feasible weights, performances, and marginal value functions. We demonstrated their usefulness on a didactic example concerning hospitals and real-world data related to the functioning of fourteen Special Economic Zones in Poland. We emphasize that the introduced framework increases the discrimination power, indicates the overall good and bad performers, and verifies the stability of efficiency results, providing hints on the required improvements and benchmarks to be followed stepwise.

We envisage the following future developments. First, the proposed framework can be extended to handle the hierarchical structure of inputs and outputs. Then, the preference information can be provided at different hierarchy levels, and the outputs can be analyzed when accounting for sub-problems, capturing specific, more consistent perspectives within a complex setting. Second, an additive value model can be revised to handle interactions between various factors. They can imply bonuses or penalties depending on whether the simultaneous favorable performances on a subset of inputs and/or outputs are exceptional or expected. Third, the scientific literature considers other imprecise performances than ordinal and interval. In particular, it would be possible to account for fuzzy performances or reasoning with evidence. Finally, it is desired to couple the proposed framework with the techniques for eliciting the shape of marginal value

functions. In this paper, we used direct elicitation procedures. However, a growing trend in MCDA is to infer such functions from the user's holistic judgments, such as pairwise comparisons.

Acknowledgments

Anna Labijak-Kowalska is grateful for the support from the Polish Ministry of Education and Science (grant no. SBAD). Miłosz Kadziński acknowledges support from the Polish National Science Center under the SONATA BIS project (grant no. DEC-2019/34/E/HS4/00045). Luis Dias acknowledges CeBER's funding by national funds through FCT – Fundação para a Ciência e a Tecnologia, Project UIDB/05037/2020.

Declarations of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- [1] Aldamak, A., Zolfaghari, S., 2017. Review of efficiency ranking methods in data envelopment analysis. *Measurement* 106, 161 – 172.
- [2] Allen, R., Athanassopoulos, A., Dyson, R. G., Thanassoulis, E., 1997. Weights restrictions and value judgements in data envelopment analysis: Evolution, development and future directions. *Annals of Operations Research* 73, 13–34.
- [3] Alvarez, P. A., Valdez, C., Dutta, B., 2022. Analysis of the innovation capacity of mexican regions with the multiple criteria hierarchy process. *Socio-Economic Planning Sciences* 84, 101418.
- [4] Bagherikahvarin, M., De Smet, Y., 2016. A ranking method based on DEA and PROMETHEE II (a rank based on DEA & PR.II). *Measurement* 89, 333–342.
- [5] Belton, V., Stewart, T. J., 1999. *DEA and MCDA: Competing or Complementary Approaches?* Springer Netherlands, Dordrecht, pp. 87–104.
- [6] Belton, V., Vickers, S. P., 1993. Demystifying dea-a visual interactive approach based on multiple criteria analysis. *The Journal of the Operational Research Society* 44 (9), 883–896.
- [7] Charnes, A., Cooper, W. W., Rhodes, E., 1978. Measuring the efficiency of decision making units. *European Journal of Operational Research* 2 (6), 429–444.
- [8] Cooper, W. W., Seiford, L. M., Tone, K., 2006. *Introduction to data envelopment analysis and its uses: with DEA-solver software and references.* Springer Science & Business Media.
- [9] de Almeida, P., Dias, L., 2012. Value-based dea models: application-driven developments. *Journal of the Operational Research Society* 63 (1), 16–27.
- [10] Fernandez, E., Figueira, J. R., Navarro, J., Solares, E., 2022. Handling imperfect information in multiple criteria decision-making through a comprehensive interval outranking approach. *Socio-Economic Planning Sciences* 82, 101254.
- [11] Gouveia, M. C., Dias, L. C., Antunes, C. H., 2008. Additive dea based on mcda with imprecise information. *Journal of the Operational Research Society* 59 (1), 54–63.
- [12] Gouveia, M. C., Dias, L. C., Antunes, C. H., 2013. Super-efficiency and stability intervals in additive DEA. *Journal of the Operational Research Society* 64 (1), 86–96.
- [13] Halme, M., Joro, T., Korhonen, P., Salo, S., Wallenius, J., 1999. A value efficiency approach to incorporating preference information in data envelopment analysis. *Management Science* 45 (1), 103–115.
- [14] Hosseinzadeh Lotfi, F., Jahanshahloo, G. R., Khodabakhshi, M., Rostamy-Malkhlifeh, M., Moghaddas, Z., Vaez-Ghasemi, M., 2013. A review of ranking models in Data Envelopment Analysis. *Journal of Applied Mathematics* 2013, 492421.
- [15] Hubinont, J., De Smet, Y., 2021. Long-term multi-criteria improvement planning. *Decision Support Systems* 149, 113606.
- [16] Joro, T., Kohonen, P., 2015. *Extension of Data Envelopment Analysis with Preference Information.* Springer, New York.
- [17] Kadziński, M., Stamenković, M., Uniejewski, M., 2022. Stepwise benchmarking for multiple criteria sorting. *Omega* 108, 102579.

- [18] Kadziński, M., Labijak, A., Napieraj, M., 2017. Integrated framework for robustness analysis using ratio-based efficiency model with application to evaluation of polish airports. *Omega* 67, 1–18.
- [19] Labijak-Kowalska, A., Kadziński, M., Spychala, I., Dias, L. C., Fiallos, J., Patrick, J., Michalowski, W., Farion, K., 2023. Performance evaluation of emergency department physicians using robust value-based additive efficiency model. *International Transactions in Operational Research* 30 (1), 503–544.
- [20] Meyer, P., Bigaret, S., 2012. Diviz: A software for modeling, processing and sharing algorithmic workflows in MCDA. *Intelligent decision technologies* 6 (4), 283–296.
- [21] Pereira, A. A., Pereira, M. A., 2023. Energy storage strategy analysis based on the Choquet multi-criteria preference aggregation model: The Portuguese case. *Socio-Economic Planning Sciences* 85, 101437.
- [22] Pereira, M. A., Figueira, J. R., Marques, R. C., 2020. Using a Choquet integral-based approach for incorporating decision-maker’s preference judgments in a Data Envelopment Analysis model. *European Journal of Operational Research* 284 (3), 1016–1030.
- [23] Petrović, M., Bojković, N., Stamenković, M., Anić, I., 2018. Supporting performance appraisal in ELECTRE based stepwise benchmarking model. *Omega* 78, 237–251.
- [24] PMET, 2018. Information about realizing the act about special economic zones (in Polish), Polish Ministry of Entrepreneurship and Technology.
- [25] Salo, A., Punkka, A., 2011. Ranking intervals and dominance relations for ratio-based efficiency analysis. *Management Science* 57 (1), 200–214.
- [26] Shen, Y., Hermans, E., Ruan, D., Wets, G., Brijs, T., Vanhoof, K., 2011. A generalized multiple layer data envelopment analysis model for hierarchical structure assessment: A case study in road safety performance evaluation. *Expert Systems with Applications* 38 (12), 15262–15272.
- [27] Stewart, T. J., 1996. Relationships between Data Envelopment Analysis and Multicriteria Decision Analysis. *Journal of the Operational Research Society* 47 (5), 654–665.
- [28] Zhu, J., 2003. Imprecise data envelopment analysis (IDEA): A review and improvement with an application. *European Journal of Operational Research* 144 (3), 513–529.

Publication [P7]

A. Labijak-Kowalska, M. Kadziński, and W. Mrozek. Robust additive value-based efficiency analysis with a hierarchical structure of inputs and outputs. *Applied Sciences*, 13(11), 2023, DOI: 10.3390/app13116406

Contribution of co-authors (excluding the author and the supervisor of this dissertation):

- Weronika Mrozek
 - Co-authorship of the idea underlying the paper,
 - reviewing and editing the text of the publication,
 - preparation of part of the software for evaluating the healthcare systems in Poland,
 - visualization of some results of the case study.

Article

Robust Additive Value-Based Efficiency Analysis with a Hierarchical Structure of Inputs and Outputs

Anna Labijak-Kowalska , Miłosz Kadziński *  and Weronika Mrozek 

Faculty of Computing and Telecommunications, Poznan University of Technology, Piotrowo 2, 60-965 Poznań, Poland; anna.labijak@cs.put.poznan.pl (A.L.-K.); weronika.mrozek@student.put.poznan.pl (W.M.)
* Correspondence: milosz.kadzinski@cs.put.poznan.pl; Tel.: +48-61-665-3022

Abstract: We introduce a novel methodological framework based on additive value-based efficiency analysis. It considers inputs and outputs organized in a hierarchical structure. Such an approach allows us to decompose the problem into manageable pieces and determine the analyzed units' strengths and weaknesses. We provide robust outcomes by analyzing all feasible weight vectors at different hierarchy levels. The analysis concerns three complementary points of view: distances to the efficient unit, ranks, and pairwise preference relations. For each of them, we determine the exact extreme results and the distribution of probabilistic results. We apply the proposed method to a case study concerning the performance of healthcare systems in sixteen Polish voivodeships (provinces). We discuss the results based on the entire set of factors (the root of the hierarchy) and three subcategories. They concern health improvement of inhabitants, efficient financial management, and consumer satisfaction. Finally, we show the practical conclusions that can be derived from the hierarchical decomposition of the problem and robustness analysis.

Keywords: data envelopment analysis; value-based efficiency; hierarchical structure; robustness analysis; healthcare



Citation: Labijak-Kowalska, A.; Kadziński, M.; Mrozek, W. Robust Additive Value-Based Efficiency Analysis with a Hierarchical Structure of Inputs and Outputs. *Appl. Sci.* **2023**, *13*, 6406. <https://doi.org/10.3390/app13116406>

Academic Editor: Konrad Kułakowski

Received: 25 April 2023

Revised: 15 May 2023

Accepted: 23 May 2023

Published: 24 May 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Data Envelopment Analysis (DEA) measures the relative efficiency of Decision Making Units (DMUs) that convert inputs to outputs. It was originally proposed by Charnes et al. [1] as a nonparametric approach, making no assumptions about the production frontier or the weights assigned to the various factors relevant to the analysis. To assess the efficiency of DMUs, they are compared to the best-practice frontier determined by the group of units with the most favorable input–output performance. The traditional methods divide the units into efficient ones, i.e., those on the efficient frontier, and inefficient ones, i.e., those below the frontier. Due to its versatility, DEA has been widely used in various areas such as management, economics, agriculture, education, healthcare, and logistics [2]. The recent example applications concerned the assessment of public administration [3] and the urban rail transit network [4].

Since its first formulation, DEA has been extended in multiple ways [5,6]. For example, various efficiency models have been introduced to admit static or dynamic analysis or handle constant or variable returns to scale. In particular, an additive model was formulated to guarantee that the units it indicates as efficient satisfy this property in Koopman's sense [7]. However, this model has also been criticized for assuming equal weights of all factors, vulnerability to the factors' scale differences, and nonintuitive interpretation of the efficiency scores. These drawbacks have motivated the development of an additive value-based efficiency analysis [8,9], inspired by Multi-Attribute Value Theory (MAVT) [10]. This model transforms the input and output values using the marginal functions. Such per-criterion components are aggregated into a comprehensive efficiency measure with an additive value model incorporating weights assigned to various factors. The units that

attain the greatest comprehensive value for at least one feasible weight vector are deemed efficient. Such an analysis is insensitive to scale problems due to applying value functions with a common scale. Moreover, the efficiency scores have an intuitive interpretation built on the notion of “min–max regret”. Note that the hybrid methods combining ideas from DEA and Multiple Criteria Decision Analysis (MCDA) have become more and more popular in recent years (see, e.g., [11,12]).

This paper contributes to the literature concerning an additive value-based efficiency analysis in a three-fold way. This methodology handles only flat structures of inputs and outputs considered at the same level, without subcategories [8]. Hence, our first contribution consists of adjusting it to handle hierarchical structures of factors used to assess the performance of DMUs. This is useful in real-world decision analysis for a few reasons. First, it helps to structure inputs and outputs logically and systematically. The higher-level factors are more general, whereas those at lower hierarchy levels are more specific. Moreover, when new information becomes available, the hierarchy can be easily modified or updated, allowing it to handle evolving decision problems. Second, a hierarchical decomposition of factors allows for the breaking down of complex problems into manageable, coherent pieces representing different levels of abstraction. By analyzing the efficiency at various levels of the hierarchy, it is possible to understand the strengths and weaknesses of DMUs and explain the comprehensive results taking into account their evolution along the hierarchy. Third, a hierarchical structure of factors makes efficiency analysis more transparent, flexible, and adaptable. In particular, we support the trade-off analysis, where weights can be associated with lower and higher-level categories of factors, and hence, their relative and absolute impact can be controlled more easily. In this regard, we incorporate the preferences elicited at each hierarchy level into the analysis. These preferences form the linear weight restrictions between factor categories at the same level.

The benefits of using a hierarchical structure have been explored in MCDA. The example methods that handle such a decomposition include the Analytical Hierarchy Process (AHP) [13], the Multiple Criteria Hierarchy Process (MCHP) [14,15], and ELECTRE-III-H [16]. In the DEA context, the first attempt was made with a two-layer nonlinear model [17] and its linear counterpart [18]. Then, Ref. [19] proposed a multiple-layer DEA model (MLDEA) handling an arbitrary number of levels of inputs and outputs. Further, MLDEA was combined with AHP to consider relative priorities of various factors, mainly in the scenarios where DEA is used as a mathematical tool for constructing so-called composite indicators [20,21]. Finally, the latter approach was generalized to the setting of Network and Fuzzy DEA [22]. The above-mentioned DEA models require inputs and outputs to be considered in separate hierarchies. We fill this research gap by admitting a single multiple-layer hierarchical structure containing inputs and outputs. In this way, the properly defined efficiency can be analyzed in each hierarchy node.

Second, in the proposed framework, we go beyond classifying the DMUs only into efficient and inefficient, as in the original value-based efficiency analysis [8]. This is attained by verifying the robustness of efficiency results observable for the entire space of feasible input and output weights. We focus on three perspectives: distances to the efficient DMU, ranks, and pairwise preference relations. For each of them, we compute the exact (necessary, possible, and extreme) outcomes by solving dedicated mathematical programming models. Moreover, we estimate the distribution of results using Monte Carlo simulations. The proposed framework is inspired by [23], being adapted to the multiple-level hierarchy value-based efficiency analysis. The formulations of dedicated procedures and types of considered results are similar to those considered in [24]. However, we admit the stakeholders studying the stability of efficiency outcomes in each hierarchy node instead of forcing them to consider all inputs and outputs simultaneously. In this regard, the main challenge is adequately handling the indicator weights considered at different hierarchy levels. We also formulate the properties of the exact efficiency outcomes observed along the hierarchy tree. Typically, they relate the results observed in all children nodes of some more

general category to the outcomes obtained in the parent node. Hence, they help understand the evolution of necessary, impossible, and extreme conclusions.

Third, we apply the proposed framework to a case study concerning healthcare. As noted in [25], the efficiency of using resources to ensure a decent level of healthcare has become one of the most critical public policy issues in recent decades. Its assessment can be conducted from the perspective of the entire system or individual organization (e.g., hospitals). A detailed review of the applications of efficiency analysis in healthcare can be found in [26], and a detailed description of the healthcare system in Poland is given in [25].

We analyzed nine indicators capturing the quality of healthcare systems in sixteen Polish voivodeships. The indicators were grouped under three main categories: inhabitants' health improvement, financial management, and consumer satisfaction. We elicited the preferences in the form of marginal value functions for all inputs and outputs and weight constraints. We report the results given a comprehensive efficiency index encompassing all relevant viewpoints and the three subproblems that allow for an understanding of each voivodeship's strong and weak points. The paper's three significant contributions, along with their essential aspects, are summarized in Figure 1.

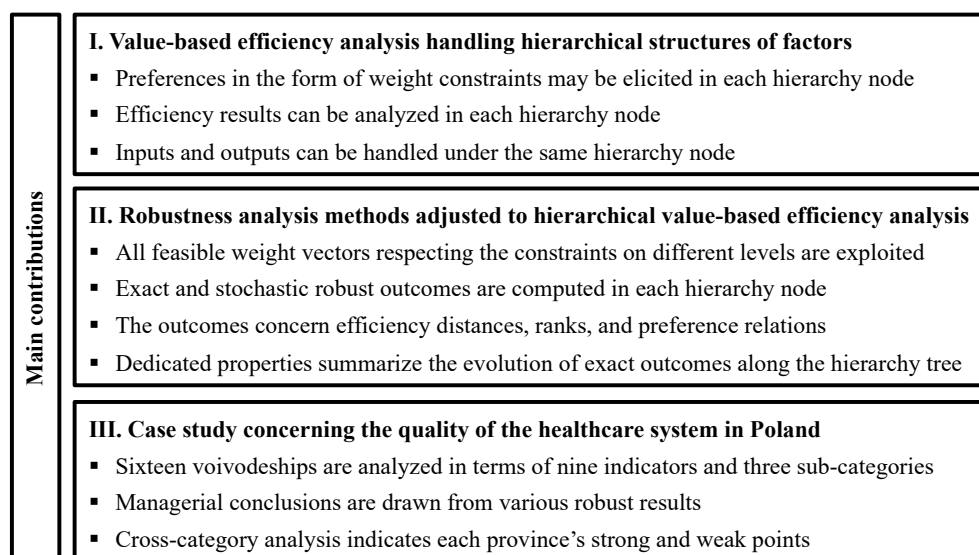


Figure 1. Paper's main contributions.

The paper's remainder is organized as follows. Section 2 describes an additive value-based efficiency model. Section 3 defines a hierarchical structure of inputs and outputs, while Section 4 describes a respective framework for robustness analysis. In Section 5, we report the outcomes of a case study concerning the efficiency assessment of the healthcare system in Poland. The last section concludes the paper.

2. Additive Value-Based Efficiency Analysis

Let us consider a set of DMUs $\mathcal{D} = \{D_1, \dots, D_K\}$. In value-based analysis, efficiency E_o of $DMU_o \in \mathcal{D}$ is defined using an additive value model:

$$E_o = \sum_{q=1}^Q w_q \cdot u_q(DMU_o), \tag{1}$$

where w_q is the weight of factor q (i.e., input $x \in \mathbf{x} = \{x_1, x_2, \dots, x_m\}$ or output $y \in \mathbf{y} = \{y_1, y_2, \dots, y_n\}$), such that $\sum_{q=1}^Q w_q = 1$, and u_q is a monotonic marginal value function for q . It is nondecreasing for outputs and nonincreasing for inputs. To verify if

DMU_o is efficient, we solve the following Linear Programming (LP) model, minimizing the maximal distance of efficiency of DMU_o to any other unit:

$$\begin{aligned}
 & \text{Minimize } d_o \\
 \text{s.t. } & \left. \begin{aligned}
 & \sum_{q=1}^Q w_q u_q(DMU_k) - \sum_{q=1}^Q w_q u_q(DMU_o) \leq d_o, \text{ for } k = 1, \dots, K, \\
 & d_o \geq 0, \\
 & \sum_{q=1}^Q w_q = 1, \\
 & w_q \geq 0, \text{ } q = 1, \dots, Q, \\
 & \mathbf{w} \in S_w,
 \end{aligned} \right\} \quad (2)
 \end{aligned}$$

where S_w is the feasible weight vector space delimited by the linear constraints. The optimal value of d_o , denoted by d_o^* , represents the minimal distance of DMU_o from the efficient unit. If $d_o^* = 0$, DMU_o attains the greatest efficiency for at least one feasible weight vector, which implies that it is efficient. When $d_o^* > 0$, for all feasible weights, there is at least one unit with an efficiency greater than E_o , denoting the lack of efficiency of DMU_o .

3. A Hierarchical Structure of Inputs and Outputs

The DMUs consume m inputs x and produce n outputs y . To simplify the notation, we aggregate all inputs and outputs into a single set of factors $f = x \cup y = \{f_1, f_2, \dots, f_{Q_0}\}$. Set f forms level 0 of the hierarchy. These factors are grouped into Q_1 categories of the first level, named $C^{(1)} = \{c_1^{(1)}, c_2^{(1)}, \dots, c_{Q_1}^{(1)}\}$. Analogously, the first-level categories can be grouped into second-level categories, forming a set $C^{(2)} = \{c_1^{(2)}, c_2^{(2)}, \dots, c_{Q_2}^{(2)}\}$, etc. The entire structure contains L levels. In the last (L -th) level, there is only a single category ($c_1^{(L)}$), called a *root*.

From the mathematical viewpoint, the factors and categories form a tree (see Figure 2). The set of all nodes in the tree (factors and categories) is denoted by $N = f \cup C^{(1)} \cup C^{(2)} \cup \dots \cup C^{(L)}$. For each node $t \in N \setminus \{root\}$, we define its parent $p(t)$ as a category in which it is directly contained. The set of direct children of category $c_i^{(l)}$ is marked as $ch(c_i^{(l)}) = \{t \in N : p(t) = c_i^{(l)}\}$. The set of indirect children of category $c_i^{(l)}$ contains all direct children of $c_i^{(l)}$ ($ch(c_i^{(l)})$) and their direct and indirect children until reaching the tree's leaves. For each category at hierarchy level $c_i^{(l)}$, we define set $A_{c_i^{(l)}}$ as a subset of f (inputs and outputs), which are the indirect children of $c_i^{(l)}$. In particular, all factors are indirect children of the root category, i.e., $A_{c_1^{(L)}} = f$. On the contrary, for an elementary factor f , A_f is a singleton, i.e., $A_f = \{f\}, f \in f$. To maintain the spirit of DEA, for each category $c_i^{(l)}$, the respective set of factors ($A_{c_i^{(l)}}$) needs to contain at least one input and one output, i.e., $A_{c_i^{(l)}} \cap x \neq \emptyset$ and $A_{c_i^{(l)}} \cap y \neq \emptyset$, for $l = 1, 2, \dots, L, i = 1, 2, \dots, Q_l$.

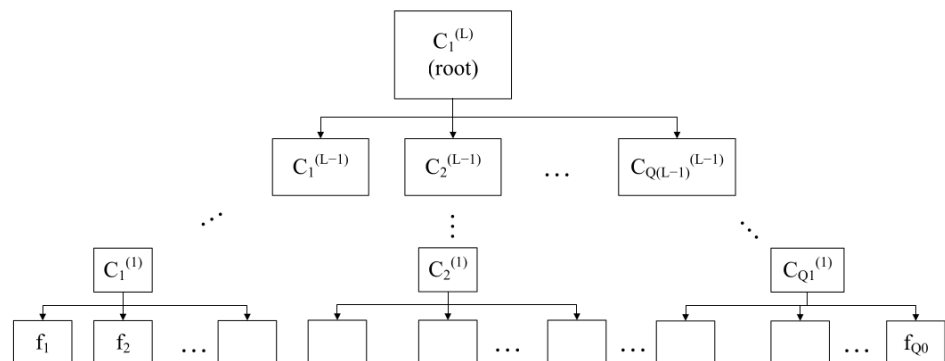


Figure 2. A hierarchical structure of inputs and outputs.

To illustrate the notation used in the paper, we will describe it using a simple hierarchy of factors in Figure 3. This example involves two inputs ($x = \{i_1, i_2\}$) and two outputs ($y = \{o_1, o_2\}$). The set of factors f containing all elements from x and y is $f = \{i_1, o_1, i_2, o_2\}$. Overall, there are four factors ($Q_0 = 4$), two first-level categories ($C_1 = \{C_1^{(1)}, C_2^{(1)}\}$; $Q_1 = 2$), and one root corresponding to a second-level category ($C_2 = \{C_1^{(2)}\}$; $Q_2 = 1$). The hierarchy contains two levels of categories ($L = 2$). The parent category for input i_1 is $C_1^{(1)}$ ($p(i_1) = C_1^{(1)}$). For $C_1^{(1)}$ and $C_2^{(1)}$, the parent category is $C_1^{(2)}$, i.e., $p(C_1^{(1)}) = p(C_2^{(1)}) = C_1^{(2)}$. The considered sets of factors $A_{C_i^{(k)}}$ for example categories are the following: $A_{i_1} = \{i_1\}$, $A_{C_1^{(1)}} = \{i_1, o_1\}$, $A_{C_2^{(1)}} = \{i_2, o_2\}$, and $A_{C_1^{(2)}} = f = \{i_1, i_2, o_1, o_2\}$.

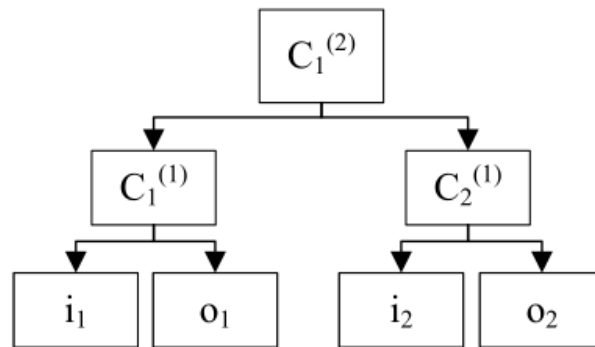


Figure 3. An example hierarchical structure of inputs and outputs.

Given a hierarchy of relevant factors and categories, we assign weight w_t to each node t except the root. Moreover, we admit specifying the linear constraints for these weights at each hierarchy level. Factors or categories involved in a single constraint must have a common parent. For example, for the considered hierarchy, the constraint can take the form $w_{i_1} \leq 2 \cdot w_{o_1}$ or $w_{C_1^{(1)}} \geq w_{C_2^{(1)}}$, as i_1 and o_1 or $C_1^{(1)}$ and $C_2^{(1)}$ have the same parent. On the contrary, the example constraint $w_{i_1} \leq w_{i_2}$ is not allowed, because i_1 and i_2 have different parents ($p(i_1) \neq p(i_2)$). The space of weight vectors that meet these restrictions is denoted by S_w .

To introduce weight restrictions, we consider additional variables (\hat{w}_q), representing the aggregated weights of elementary factors in the hierarchy. They are defined as the products of all weights on the path from the analyzed category ($c_i^{(l)}$) at the hierarchy level l to the analyzed factor f :

$$\hat{w}_q^{c_i^{(l)}} = w_q \cdot \prod_{t=1, \dots, l-1 \wedge t \in c_i^{(l)} \wedge f \in A_t} w_t. \tag{3}$$

For factor i_1 in the considered example, when taking into account the root category ($C_1^{(2)}$), the above formula takes the following form $\hat{w}_{i_1}^{C_1^{(2)}} = w_{i_1} \cdot w_{C_1^{(1)}}$, and when considering category $C_1^{(1)}$, it is expressed as follows $\hat{w}_{i_1}^{C_1^{(1)}} = w_{i_1}$.

We analyze the efficiency of DMU_o in each node of the hierarchy. For category $c_i^{(l)}$, such efficiency is defined as follows:

$$E_o^{c_i^{(l)}} = \sum_{q \in A_{c_i^{(l)}}} \hat{w}_q^{c_i^{(l)}} \cdot u_q(DMU_o). \tag{4}$$

The true weights assigned to each hierarchy category $c_i^{(l)}$ from the set of the indirect children of the analyzed category $c_j^{(k)}$ are defined as the ratio of the sum of weights of indicators contained in this category and the sum of weights of indicators in the parent category:

$$w_{c_i^{(l)}}^{c_j^{(k)}} = \frac{\sum_{f \in A_{c_i^{(l)}}} \hat{w}_f^{c_j^{(k)}}}{\sum_{f_p \in A_{p(c_i^{(l)})}} \hat{w}_{f_p}^{c_j^{(k)}}}. \tag{5}$$

Note that the value of weight $w_{c_i^{(l)}}^{c_j^{(k)}}$ is always the same, regardless of the considered category ($c_j^{(k)}$), so we replace symbol $w_{c_i^{(l)}}^{c_j^{(k)}}$ with $w_{c_i^{(l)}}$. For example, when considering the root category ($C_1^{(2)}$), the weight w_{i_1} of indicator i_1 in the considered example can be calculated as $w_{i_1} = \frac{\hat{w}_{i_1}^{C_1^{(2)}}}{\hat{w}_{i_1}^{C_1^{(2)}} + \hat{w}_{o_1}^{C_1^{(2)}}$, whereas the weight $w_{C_2^{(1)}}$ of category $C_2^{(1)}$ is

$$w_{C_2^{(1)}} = \frac{\hat{w}_{i_2}^{C_1^{(2)}} + \hat{w}_{o_2}^{C_1^{(2)}}}{\hat{w}_{i_1}^{C_1^{(2)}} + \hat{w}_{o_1}^{C_1^{(2)}} + \hat{w}_{i_2}^{C_1^{(2)}} + \hat{w}_{o_2}^{C_1^{(2)}}}.$$

4. Robustness Analysis for Additive Value-Based Efficiency Analysis with a Hierarchical Structure of Factors

The standard value-based efficiency model verifies if each DMU is efficient. Such an analysis builds on the weight vector that is the most advantageous for a given DMU, allowing it to minimize the distance from some efficient DMU. In this section, we introduce a suite of methods that investigate the robustness of efficiency outcomes given all feasible weights. They can be divided into two groups. First, the exact approaches use mathematical programming to find the extreme outcomes for each DMU. In turn, the probabilistic methods estimate the stochastic acceptability indices based on Monte Carlo simulations, reflecting the distributions of possible results. Each group concerns three relevant viewpoints: distances to the efficient unit, ranks, and pairwise preference relations. In what follows, we present the approaches that are flexible enough to determine the relevant results in each hierarchy node.

4.1. Exact Methods

In this section, we present the mathematical programming models that determine the exact robust results. These include extreme (the most and the least advantageous), necessary (observable for all feasible weight vectors), and possible (holding for at least one feasible weight vector) conclusions. Let us first focus on verifying the stability of distances to the efficient unit. The best (minimal) distance $d_{*,o}^{c_i^{(l)}}$ for DMU_o considering category $c_i^{(l)}$ can be computed by solving the following model:

$$\text{Minimize } d_o^{c_i^{(l)}}$$

s.t.

$$\left. \begin{aligned} & \sum_{q \in A_{c_i^{(l)}}} \hat{w}_q^{c_i^{(l)}} u_q(DMU_k) - \sum_{q \in A_{c_i^{(l)}}} \hat{w}_q^{c_i^{(l)}} u_q(DMU_o) \leq d_o^{c_i^{(l)}}, \text{ for } k = 1, \dots, K, \\ & d_o^{c_i^{(l)}} \geq 0, \\ & \sum_{q \in A_{c_i^{(l)}}} \hat{w}_q^{c_i^{(l)}} = 1, \\ & \hat{w}_q \geq 0, \text{ } q \in A_{c_i^{(l)}}, \\ & w_t = \frac{\sum_{f \in A_t} \hat{w}_f^{c_i^{(l)}}}{\sum_{f \in A_{p(t)}} \hat{w}_f^{c_i^{(l)}}} \in S_w, \text{ for } t \in N \setminus \{\text{root}\}. \end{aligned} \right\} \mathcal{W} \tag{6}$$

Similarly to the standard efficiency analysis, DMU_o with $d_{*,o}^{c_i^{(l)}} = 0$ is deemed efficient, given category $c_i^{(l)}$, while $d_{*,o}^{c_i^{(l)}} > 0$ implies inefficiency.

To compute the worst (maximal) distance $d_o^{*,c_i^{(l)}}$ for DMU_o , given category $c_i^{(l)}$, we solve the following Mixed-Integer Linear Programming (MILP) model:

$$\begin{aligned} & \text{Maximize } d_o^{c_i^{(l)}} \\ & \text{s.t.} \\ & \left. \begin{aligned} & \sum_{q \in A_{c_i^{(l)}}} \hat{w}_q^{c_i^{(l)}} u_q(DMU_k) - d_o^{c_i^{(l)}} \geq \sum_{q \in A_{c_i^{(l)}}} \hat{w}_q^{c_i^{(l)}} u_q(DMU_o) - C(1 - b_k), \text{ for } k = 1, \dots, K, \\ & \sum_{k=1, \dots, K} b_k = 1, \\ & b_k \in \{0, 1\}, \text{ for } k = 1, 2, \dots, K, \\ & d_o^{c_i^{(l)}} \geq 0, \\ & \mathcal{W}, \end{aligned} \right\} \tag{7} \end{aligned}$$

where C is a large positive constant. The above model uses binary variables b_k , $k = 1, \dots, K$, to ensure that $d_o^{c_i^{(l)}}$ is equal to the efficiency difference between DMU_o and some DMU_k , $k = 1, \dots, K$, for which $b_k = 1$. Maximizing $d_o^{c_i^{(l)}}$ guarantees that we obtain the greatest possible difference observable in the set of feasible weights \mathcal{W} . Note that when $b_k = 0$, the respective constraint is satisfied for all possible variable values; hence, it is relaxed.

The second perspective concerns the bounds of efficiency ranks attained by DMU_o . To find the best (minimal) rank $R_{*,o}^{c_i^{(l)}}$ of DMU_o , given category $c_i^{(l)}$, we minimize the number of other DMUs with greater efficiencies than $E_o^{c_i^{(l)}}$:

$$\begin{aligned} & \text{Minimize } 1 + \sum_{k=1, \dots, K; k \neq o} b_k \\ & \text{s.t.} \\ & \left. \begin{aligned} & \sum_{q \in A_{c_i^{(l)}}} \hat{w}_q^{c_i^{(l)}} u_q(DMU_k) - \sum_{q \in A_{c_i^{(l)}}} \hat{w}_q^{c_i^{(l)}} u_q(DMU_o) \leq C b_k, \text{ for } k = 1, \dots, K; k \neq o, \\ & b_k \in \{0, 1\}, \text{ for } k = 1, 2, \dots, K; k \neq o, \\ & \mathcal{W}. \end{aligned} \right\} \tag{8} \end{aligned}$$

Note that when $b_k = 0$, $k = 1, \dots, K$, the respective constraint ensures that DMU_o is ranked not worse than DMU_k since $E_k^{c_i^{(l)}} \leq E_o^{c_i^{(l)}}$. When $b_k = 1$, DMU_k is ranked better than DMU_o , deteriorating its best rank by one.

To obtain the worst (maximal) rank $R_o^{*,c_i^{(l)}}$ for DMU_o , given category $c_i^{(l)}$, we maximize the number of DMUs with the efficiencies not worse than $E_o^{c_i^{(l)}}$:

$$\text{Maximize } 1 + \sum_{k=1, \dots, K; k \neq o} b_k$$

s.t.

$$\left. \begin{aligned} &\sum_{q \in A_{c_i^{(l)}}} \hat{w}_q^{c_i^{(l)}} u_q(DMU_o) - \sum_{q \in A_{c_i^{(l)}}} \hat{w}_q^{c_i^{(l)}} u_q(DMU_k) \leq C(1 - b_k), \text{ for } k = 1, \dots, K; k \neq o, \\ &b_k \in \{0, 1\}, \text{ for } k = 1, 2, \dots, K; k \neq o, \\ &\mathcal{W}. \end{aligned} \right\} \quad (9)$$

Note that when $b_k = 1, k = 1, \dots, K$, the respective constraint ensures that DMU_k is ranked no worse than DMU_o since $E_o^{c_i^{(l)}} \leq E_k^{c_i^{(l)}}$. This deteriorates the worst rank of DMU_o by one. When $b_k = 0$, the respective constraint is satisfied for all variable values; hence, it relaxed.

The third perspective focuses on the pairwise comparisons between DMUs using two relations: necessary and possible. Given the uncertainty of selecting a specific weight vector, the necessary relation can be considered robust. Specifically, DMU_o is necessarily preferred to DMU_k , given category $c_i^{(l)}$ ($DMU_o \succsim_E^{N,c_i^{(l)}} DMU_k$), when DMU_o is not worse at level $c_i^{(l)}$ in terms of efficiency than DMU_k for all feasible weight vectors. Its truth for pair (DMU_o, DMU_k) and category $c_i^{(l)}$ can be verified using the following model:

$$\text{Minimize } d_{o,k}^{c_i^{(l)}}, \text{ s.t. } \sum_{q \in A_{c_i^{(l)}}} \hat{w}_q^{c_i^{(l)}} u_q(DMU_o) - \sum_{q \in A_{c_i^{(l)}}} \hat{w}_q^{c_i^{(l)}} u_q(DMU_k) \leq d_{o,k}^{c_i^{(l)}} \text{ and } \mathcal{W}. \quad (10)$$

Its optimal solution $d_{o,k}^{c_i^{(l)},*}$ is equal to the minimal difference between efficiencies of DMU_o and DMU_k observable in the set of feasible weights \mathcal{W} , given category $c_i^{(l)}$. If $d_{o,k}^{c_i^{(l)},*} \geq 0$, then for all feasible weights $E_o^{c_i^{(l)}} \geq E_k^{c_i^{(l)}}$, and hence $DMU_o \succsim_E^{N,c_i^{(l)}} DMU_k$. Otherwise, $\neg(DMU_o \succsim_E^{N,c_i^{(l)}} DMU_k)$ because there is at least one feasible weight vector, such that $E_o^{c_i^{(l)}} < E_k^{c_i^{(l)}}$.

Furthermore, DMU_o is possibly preferred to DMU_k , given category $c_i^{(l)}$ ($DMU_o \succsim_E^{P,c_i^{(l)}} DMU_k$), when DMU_o is not worse at level $c_i^{(l)}$ in terms of efficiency than DMU_k for at least one feasible weight vector. Its truth for pair (DMU_o, DMU_k) and category $c_i^{(l)}$ is verified using the following model:

$$\text{Maximize } d_{o,k}^{c_i^{(l)}}, \text{ s.t. } \sum_{q \in A_{c_i^{(l)}}} \hat{w}_q^{c_i^{(l)}} u_q(DMU_o) - \sum_{q \in A_{c_i^{(l)}}} \hat{w}_q^{c_i^{(l)}} u_q(DMU_k) \geq d_{o,k}^{c_i^{(l)}} \text{ and } \mathcal{W}. \quad (11)$$

Its optimal solution $d_{o,k}^{c_i^{(l)},*}$ is equal to the maximal difference between efficiencies of DMU_o and DMU_k observable in the set of feasible weights \mathcal{W} , given category $c_i^{(l)}$. If $d_{o,k}^{c_i^{(l)},*} \geq 0$, then for at least one feasible weight $E_o^{c_i^{(l)}} \geq E_k^{c_i^{(l)}}$, and hence, $DMU_o \succsim_E^{P,c_i^{(l)}} DMU_k$. Otherwise, $\neg(DMU_o \succsim_E^{P,c_i^{(l)}} DMU_k)$ because there is no feasible weight vector, such that $E_o^{c_i^{(l)}} \geq E_k^{c_i^{(l)}}$.

The relevant properties of the exact robust results given the hierarchical structure are presented in Appendix A. The formulations of example mathematical programming models that support understanding the general formulations are given in Appendix B.

4.2. Simulation-Based Methods

The results determined with mathematical programming are often insufficiently conclusive. In particular, the difference between extreme distances or ranks may be significant, the necessary relation may be poor, and the possible relation may be very rich. If so, it would be helpful to determine the distribution of results observed for the set of feasible weight vectors. Unfortunately, such distribution cannot be computed exactly. However, using Monte Carlo simulations, we can estimate the share of feasible weight space confirming a particular outcome. Specifically, we use the hit-And-run algorithm to generate a predefined number of weight vector samples [27]. We generate weights for all categories and factors while respecting that the sum of weights of categories or factors with the same parent must be equal to one. In the example problem, the sum of the weights of two categories $C_1^{(1)}$ and $C_2^{(1)}$ must be equal to one ($w_{C_1^{(1)}} + w_{C_2^{(1)}} = 1$), and the sums of weights assigned to the elementary indicators in the same category also need to be one ($w_{i_1} + w_{o_1} = 1$ and $w_{i_2} + w_{o_2} = 1$). Moreover, we obey the provided weight constraints for all hierarchy levels. After generating a predefined number of weight samples, we compute the efficiencies of all DMUs for each. This lets us calculate the relevant stochastic acceptability indices estimating the respective shares of feasible weight vectors.

In what follows, when considering category $c_i^{(l)}$ and referring to weight vectors, we mean the weights assigned to all categories and factors that are direct or indirect children of $c_i^{(l)}$ in the hierarchy. The most interesting stochastic acceptabilities are defined as follows:

- Distance Acceptability Interval Index ($DAII^{c_i^{(l)}}(DMU_o, b_i)$) for unit DMU_o , interval $b_i \subseteq [0, 1]$, and category $c_i^{(l)}$ is the share of feasible weight vectors for which $E_o^{c_i^{(l)}}$ belongs to b_i . Note that all intervals must be disjoint ($b_i \cap b_j = \emptyset, i \neq j$), and their sum must cover the space of possible distances ($b_1 \cup b_2 \cup \dots \cup b_z = [0, 1]$; z —the number of intervals).
- Efficiency Rank Acceptability Index ($ERAI^{c_i^{(l)}}(DMU_o, r)$) for unit DMU_o and rank r is the share of feasible weight vectors for which DMU_o attains r -th position in the efficiency ranking of all DMUs given category $c_i^{(l)}$.
- Pairwise Efficiency Outranking Index ($PEOI^{c_i^{(l)}}(DMU_o, DMU_k)$) for pair (DMU_o, DMU_k) and category $c_i^{(l)}$ is the share of feasible weight vectors for which DMU_o is at least as efficiency as DMU_k at level $c_i^{(l)}$, i.e., $E_o^{c_i^{(l)}} \geq E_k^{c_i^{(l)}}$.

Moreover, we compute the expected distance Ed to the efficient unit and expected rank ER for each DMU [28]. This is performed by averaging the distances or ranks observed for all samples. Note that by default, we use uniform distribution for weight sampling. However, the weights can be generated from any predefined distribution, but it is hard to define as it requires in-depth knowledge about the specific application domain.

In Appendix C, we illustrate the process of computing the stochastic results on a small sample of weight vectors.

5. Case Study concerning Evaluation of Healthcare System in Poland

This section reports the results of a case study concerning an assessment of the quality of the healthcare system in Poland. This sector faces the challenge of improving the quality of provided services. This can be attained by advancing some indicators reflecting both the system's functioning and the perception by patients. We consider sixteen voivodeships (provinces) in Poland as DMUs (see Table 1). These administrative areas govern their healthcare independently, so it makes sense to highlight their differences using a uniformly

computed set of indicators. Such an evaluation is critical, given the rapid development of new technologies and major transformations in the healthcare sector.

Table 1. Values of inputs and outputs related to the healthcare systems in sixteen Polish voivodeships [29] (data publicly available at <https://www.pwc.pl/pl/publikacje/2019/indeks-sprawnosci-ochrony-zdrowia-2018.html>, accessed on 22 May 2023).

Voivodeship	Short Name	H_1	H_2	H_3	F_1	F_2	F_3	S_1	S_2	S_3
Zachodniopomorskie	ZPM	44.44	4.05	17.6	−1.5	46.1	46.9	19.23	3.55	9.30
Pomorskie	POM	45.31	5.36	21.3	0.08	38.9	51.3	27.04	3.82	33.33
Warmińsko–Mazurskie	WM	43.34	7.11	15.7	−1.09	44.8	47.8	22.19	3.72	22.73
Podlaskie	PDL	37.54	6.93	15.8	2.3	49.4	43.4	19.47	3.61	11.11
Lubuskie	LBU	50.21	5.83	18.0	1.9	42.0	51.3	18.32	3.84	13.04
Wielkopolskie	WLKP	47.88	4.36	25.7	−0.3	43.4	50.5	15.16	3.69	17.24
Kujawsko–Pomorskie	KP	39.90	6.80	15.7	3.8	46.5	41.8	22.36	3.77	33.33
Mazowieckie	MAZ	38.59	5.32	16.7	−1.57	47.4	47.8	20.26	3.62	20.56
Lubelskie	LBL	45.17	6.74	8.1	0.41	51.8	43.0	13.40	3.77	20.37
Dolnośląskie	DSL	45.04	7.66	13.8	−1.77	50.8	43.8	24.55	3.57	17.28
Opolskie	OPO	33.71	6.88	10.0	−0.36	44.1	40.9	23.03	3.67	35.71
Łódzkie	LDZ	42.00	5.88	6.0	1.02	50.6	47.3	19.90	3.78	15.15
Śląskie	SL	40.70	4.54	13.9	−1.63	55.5	39.1	22.02	3.75	16.88
Świętokrzyskie	SW	41.14	5.79	7.8	0.17	49.5	47.2	12.44	3.57	29.17
Małopolskie	MLP	32.22	5.61	9.1	0.79	43.7	44.3	18.27	3.64	25.84
Podkarpackie	PKR	38.73	5.76	1.9	−4.31	48.0	45.8	14.39	3.84	20.51

Following [29], we consider three main categories of factors representing desirable characteristics in the complex healthcare system. Two areas—health improvement and financial management—are based on objective indicators and parameters, whereas the system’s evaluation by patients is, to some extent, subjective. In particular, improving health is the ultimate aim of the healthcare system. In this regard, it is relevant to consider the example dimensions of the health status that are affected by how the system is operated, given the ever-growing needs of patients. Financial management is essential in healthcare, as this sector experiences the availability of limited resources. Hence, it is vital to assess the financial situation of medical facilities, the management of infrastructure, and the economic efficiency of treatments and therapies. Finally, consumer satisfaction is becoming more and more important in evaluating the healthcare sector. Thus, it is desirable to consider the quality of services, comfort in using patients’ services, and patient rights.

The hierarchy of inputs and outputs for the case study is presented in Figure 4. Among the nine factors, there are six outputs (H_1 , H_3 , F_1 , F_3 , S_2 , and S_3) and three inputs (H_2 , F_2 , and S_1). The selected indicators are representative of the three dimensions and the viewpoints of the most important stakeholders. The indicators in the health improvement category (H) are representative of the dimensions of preventing diseases (H_1), their exacerbation (H_2), and deaths (H_3). The factors considered in the financial management category stand for the financial situation of healthcare units (F_1) and infrastructure management (F_2 and F_3). The inputs and outputs in the system’s evaluation category (S) represent the waiting time (S_1), official quality system (S_2), and patient satisfaction (S_3). Moreover, we verified that the trends that these indicators confirm also represent other factors that could be considered in the three categories. Note that analysis including over 40 indicators available for assessing the healthcare system in Poland [29] would not make much sense in the context of DEA, as the number of inputs and outputs would be too large compared to the number of DMUs. Typically, such analyses indicate that all or almost all units are efficient, as even the worst performers tend to specialize in some particular aspects. Hence, we opted for an analysis with a reduced—though carefully selected—set of indicators. The performances of the sixteen voivodeships in terms of nine considered factors are given in Table 1.

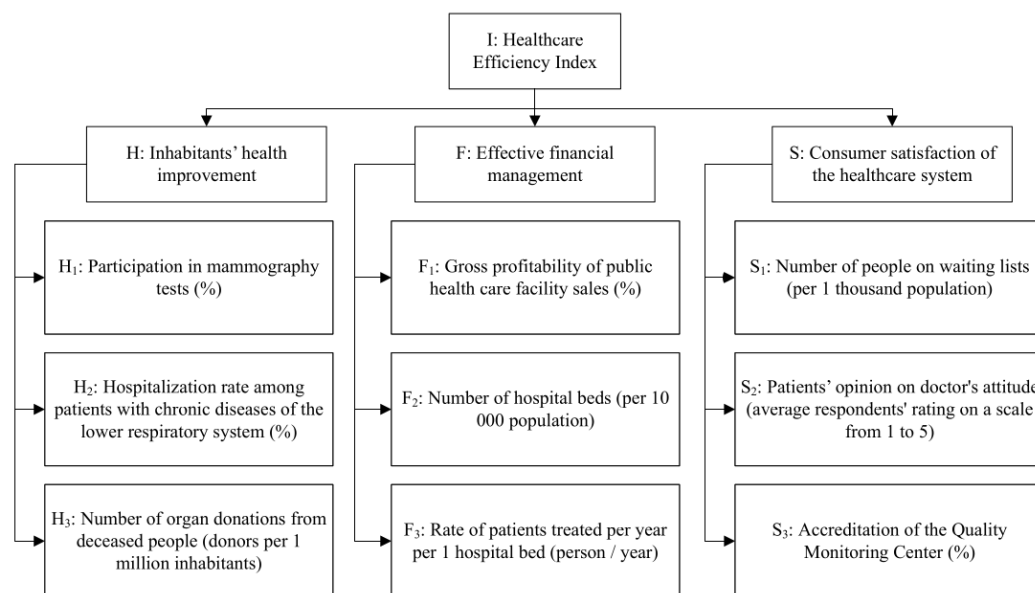


Figure 4. A hierarchy of indicators considered in the quality evaluation of healthcare systems.

For all factors, we elicited marginal value functions from experts in the healthcare system in Poland. They are provided in Figure 5. They are decreasing for inputs and increasing for outputs. Moreover, they differ in shape. The function is, e.g., close to linear for H_2 , convex for F_3 , concave for F_2 , and S-shaped for S_2 . Moreover, we incorporated the relative and absolute weight constraints. The category of *inhabitants' health improvement* is more important than the two other categories. Hence, we introduced the following constraints: $w_H \geq w_F$ and $w_H \geq w_S$, where w_H , w_F , and w_S are the weights of the three categories. Finally, we wanted to avoid both the minor and dominating roles of any individual factor or category in the analysis. Hence, we restricted the weight of each category to be not less than 0.2 and the weights of second-level elementary factors to the interval $[0.2, 0.5]$. In what follows, we discuss the results attained in the root hierarchy level and for each of the three categories separately.

5.1. Comprehensive Evaluation of the Quality of Healthcare Systems

In this section, we discuss the results of the comprehensive assessment of Polish voivodeships, taking into account all nine indicators. Figure 6 presents the extreme and expected distances to the best unit for each analyzed province. Three voivodeships are efficient: POM, LBU, and WLKP. POM also attains the lowest maximal distance (0.105), which confirms its most favorable evaluation of the healthcare system for all feasible weights. Moreover, the distances for POM are the most stable, as it is characterized by the narrowest range ($d^{l,*} - d_*^l = 0.105$). Among the efficient provinces, WLKP has the worst pessimistic distance to the best province. However, its expected distance is better than that of LBU, meaning that for some weight vectors, WLKP performs worse than LBU, but its efficiency score is closer to the best province on average. The worst provinces in the most and the least favorable scenarios are DSL ($d_*^l = 0.306$, $d^{l,*} = 0.522$) and SL ($d_*^l = 0.251$, $d^{l,*} = 0.526$). The greatest sensitivity of the distances depending on the selected weight vector is observed for SW, as the width of its distance interval equals 0.355.

The analysis of extreme distances can be enriched with the distribution of distances over all feasible weight vectors (see Table 2). The efficient provinces (POM, LBU, and WLKP) are the only ones whose distances were not greater than 0.1 for some samples. When considering these three voivodeships, only for LBU, the distance was greater than 0.1 for some marginal share of weights (0.8%). Hence, these provinces are robustly better than the remaining ones. Among the inefficient units, the most favorable results were attained by ZPM and KP.

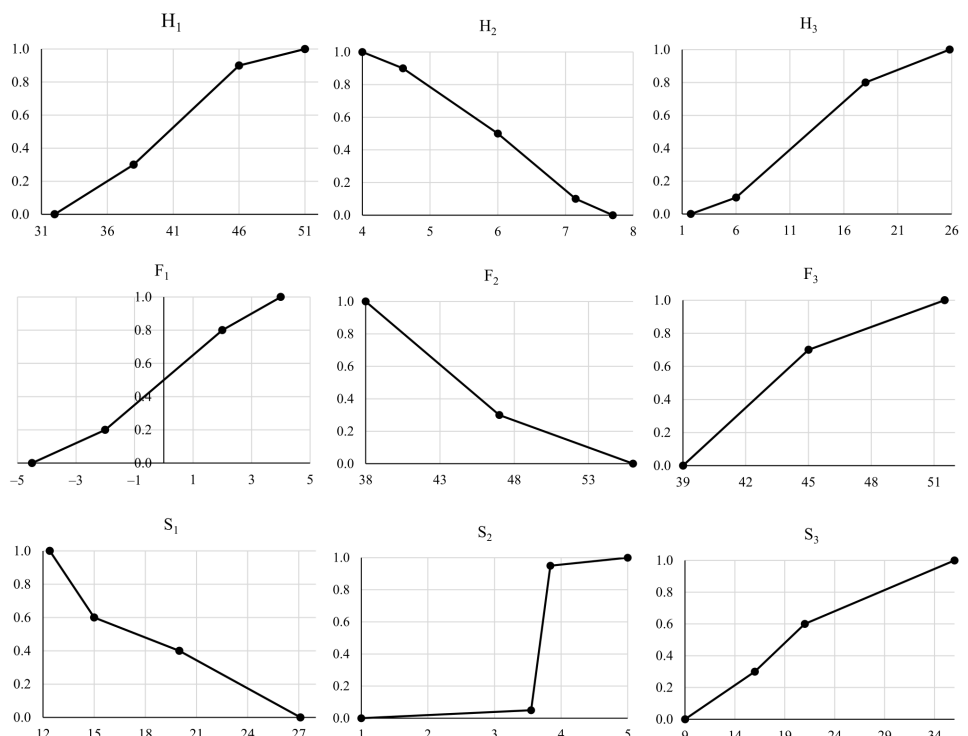


Figure 5. The marginal value functions associated with the inputs and outputs considered in the case study.

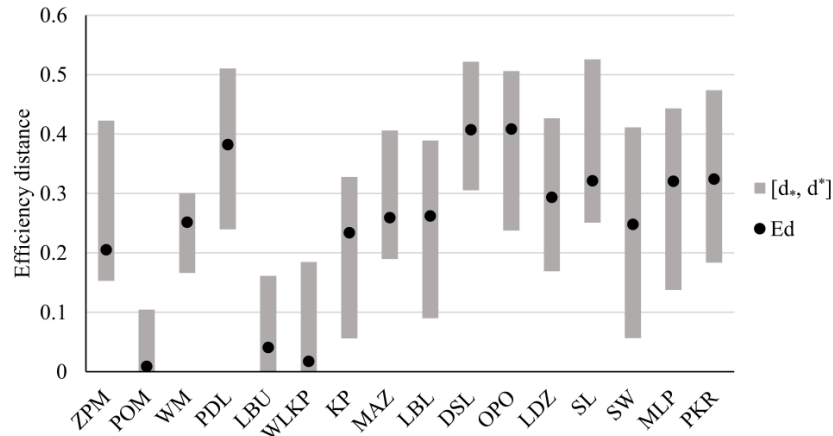


Figure 6. Extreme and expected distances to the best unit for Polish voivodeships in the comprehensive analysis of the healthcare system.

Some provinces are characterized by rather stable distance values. For example, for MAZ and WM, for over 95% of weight vectors, the distance is the interval (0.2, 0.3]. On the contrary, the distance for PKR varies more depending on the chosen weight vector. In this case, positive *DAIIs* were observed for all buckets between 0.1 and 0.5 with the greatest values for the intervals (0.3, 0.4] ($DAII(PKR, (0.3, 0.4]) = 65.9\%$) and [0.2, 0.3] ($DAII(PKR, (0.2, 0.3]) = 29.5\%$). The complete ranking determined by the expected distances (see Figure 6) indicates POM (0.009) and WLKP (0.017) as the best units and DSL (0.408) and OPO (0.409) as the worst.

Table 2. Distribution of distances to the best unit for Polish voivodeships in the comprehensive analysis of the healthcare system.

DAII	[0.0, 0.1]	(0.1, 0.2]	(0.2, 0.3]	(0.3, 0.4]	(0.4, 0.5]	(0.5, 0.6]	(0.6, 0.7]	(0.7, 0.8]	(0.8, 0.9]	(0.9, 1.0]
ZPM	0	0.523	0.475	0.002	0	0	0	0	0	0
POM	1	0	0	0	0	0	0	0	0	0
WM	0	0.006	0.956	0.038	0	0	0	0	0	0
PDL	0	0	0.006	0.714	0.280	0	0	0	0	0
LBU	0.992	0.008	0	0	0	0	0	0	0	0
WLKP	1	0	0	0	0	0	0	0	0	0
KP	0	0.160	0.798	0.042	0	0	0	0	0	0
MAZ	0	0.001	0.963	0.036	0	0	0	0	0	0
LBL	0	0.074	0.729	0.197	0	0	0	0	0	0
DSL	0	0	0	0.406	0.594	0	0	0	0	0
OPO	0	0	0	0.407	0.587	0.006	0	0	0	0
LDZ	0	0	0.599	0.401	0	0	0	0	0	0
SL	0	0	0.254	0.732	0.014	0	0	0	0	0
SW	0	0.127	0.762	0.111	0	0	0	0	0	0
MLP	0	0	0.307	0.675	0.018	0	0	0	0	0
PKR	0	0.002	0.295	0.659	0.044	0	0	0	0	0

The results of robustness analysis for efficiency ranks are presented in Figure 7 and Table 3. The three efficient voivodeships attain the first rank in the most favorable scenario. POM and LBU are ranked third in their worst scenario, while WLKP falls fourth in the pessimistic case. These units are also the best, given their expected ranks. In this regard, POM (1.543) is followed by WLKP (1.783) and LBU (2.674). Among the inefficient units, KP is the most advantageous when considering the most favorable ranks ($R_* = 2$). Other inefficient units ranked relatively high in their best scenario are ZPM, WM, LBL, LDZ, SW, MLP, and PKR ($R_* = 4$). However, all inefficient provinces are ranked low in the least favorable scenario. The best maximal rank among them is observed for KP and LBL ($R^* = 12$), while five provinces can be ranked at the bottom (PDL, DSL, OPO, SL, and PKR). Finally, the best expected ranks among inefficient units are attained by ZPM (4.962), KP (5.878), and SW (6.611), while the worst expected positions are associated with DSL (15.166) and OPO (15.195).

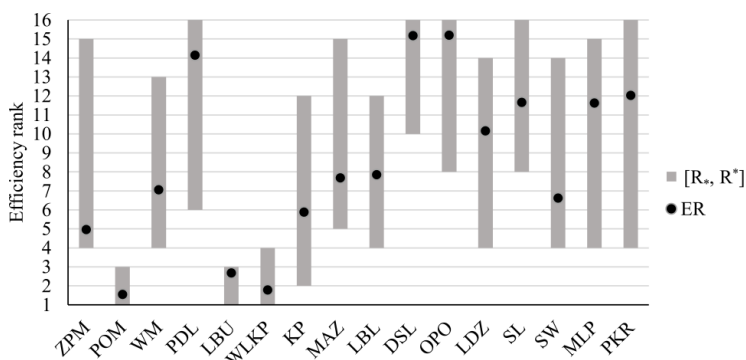


Figure 7. Extreme and expected ranks for Polish voivodeships in the comprehensive analysis of the healthcare system.

The analysis of efficiency rank acceptability indices (see Table 3) confirms the superiority of POM over other provinces. It is ranked first for most feasible weight vectors (57.7%), and it is in the top two for almost 90% of samples. Similarly, WLKP is at least second for over 84% of samples, though its most frequent position is second rather than first. In turn, LBU is ranked third for most scenarios (72.4%). Even though the best possible rank for KP is second, such a position was not observed for any weight vector. The highest for which

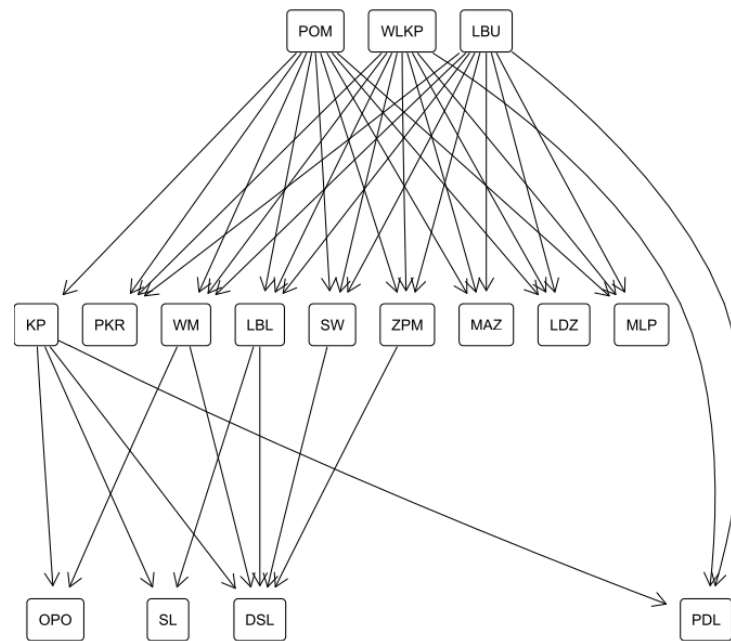


Figure 8. The Hasse diagram of the necessary efficiency preference relation in the comprehensive analysis of the healthcare system.

For pairs of voivodeships that are not related by the necessary preference, it is worth analyzing the pairwise efficiency outranking indices (see Table 5). For some, one province proves better for most scenarios (see, e.g., ZPM and PKR with $PEOI^I(ZPM, PKR) = 99.4\%$ or MAZ and MLP with $PEOI^I(MAZ, MLP) = 99.3\%$). For other pairs, the shares of feasible scenarios confirming the preference in both directions are more balanced (see, e.g., MAZ and LBL with $PEOI^I(MAZ, LBL) = 53.4\%$ and $PEOI^I(LBL, MAZ) = 46.6\%$, or PKR and SL with $PEOI^I(PKR, SL) = 47.6\%$ and $PEOI^I(SL, PKR) = 52.4\%$).

Table 5. Pairwise efficiency outranking indices for pairs of Polish voivodeships in the comprehensive analysis of the healthcare system.

<i>PEOI</i>	ZPM	POM	WM	PDL	LBU	WLKP	KP	MAZ	LBL	DSL	OPO	LDZ	SL	SW	MLP	PKR
ZPM	1	0	0.835	1	0	0	0.690	0.962	0.852	1	1	0.989	1	0.791	1	0.994
POM	1	1	1	1	0.870	0.625	1	1	1	1	1	1	1	1	1	1
WM	0.165	0	1	1	0	0	0.286	0.609	0.630	1	1	0.890	0.918	0.481	0.960	0.967
PDL	0	0	0	1	0	0	0	0	0	0.760	0.748	0.006	0.095	0	0.039	0.146
LBU	1	0.130	1	1	1	0.181	1	1	1	1	1	1	1	1	1	1
WLKP	1	0.375	1	1	0.819	1	1	1	1	1	1	1	1	1	1	1
KP	0.310	0	0.714	1	0	0	1	0.758	0.805	1	1	0.953	0.974	0.651	0.996	0.990
MAZ	0.038	0	0.391	1	0	0	0.242	1	0.534	1	1	0.850	0.981	0.378	0.993	0.958
LBL	0.148	0	0.370	1	0	0	0.195	0.466	1	1	1	0.879	0.909	0.295	0.898	0.988
DSL	0	0	0	0.240	0	0	0	0	0	1	0.487	0.001	0.016	0	0.052	0.061
OPO	0	0	0	0.252	0	0	0	0	0	0.513	1	0.002	0.023	0	0	0.020
LDZ	0.011	0	0.110	0.994	0	0	0.047	0.150	0.121	0.999	0.998	1	0.745	0.051	0.757	0.849
SL	0	0	0.082	0.905	0	0	0.026	0.019	0.091	0.984	0.977	0.255	1	0.049	0.489	0.524
SW	0.209	0	0.519	1	0	0	0.349	0.622	0.705	1	1	0.949	0.951	1	0.997	0.992
MLP	0	0	0.040	0.961	0	0	0.004	0.007	0.102	0.948	1	0.243	0.511	0.003	1	0.568
PKR	0.006	0	0.033	0.854	0	0	0.010	0.042	0.012	0.939	0.980	0.151	0.476	0.008	0.432	1

5.2. The Category of Inhabitants' Health Improvement

In this section, we focus on the results attained when considering only inputs and outputs from the inhabitants' health improvement category. In addition, we emphasize the differences with respect to the comprehensive level.

Figure 9 presents the extreme and expected distances to the best unit for all considered voivodeships. WLKP is the only efficient province given this category, so its distance to

the best unit always equals zero. Hence, POM and LBU lose the status of efficient units. However, these two voivodeships and ZPM have relatively low distances to the best one: ZPM ($d^H \in [0.065, 0.146]$), POM ($d^H \in [0.130, 0.180]$), and LBU ($d^H \in [0.114, 0.246]$). In general, the widths of distance intervals are notably more precise than when considering all relevant factors jointly. They are also more diverse, encompassing a greater range. In particular, seven provinces have maximal distances greater than 0.6, with three even exceeding the threshold of 0.7. At the comprehensive level, this was not observed for any voivodeship.

Among the inefficient provinces, ZPM can be considered the best, as it has the lowest distances to the best unit in both optimistic and pessimistic settings. Moreover, its best expected distance (0.109) is twice lower than at the comprehensive level, letting it overtake POM and LBU, which were judged efficient in the hierarchy's root. The other twelve inefficient provinces are significantly worse. The least favorable among them is OPO, with the distance to the best unit in its optimistic scenario equal to 0.73 and an expected distance of 0.754. In the average case, OPO is directly preceded by MLP (0.659) and PKR (0.652). Note that at the comprehensive level, the worst maximal distances were attained by SL, DSL, and PDL. They all prove slightly better regarding inhabitants' health improvement results.

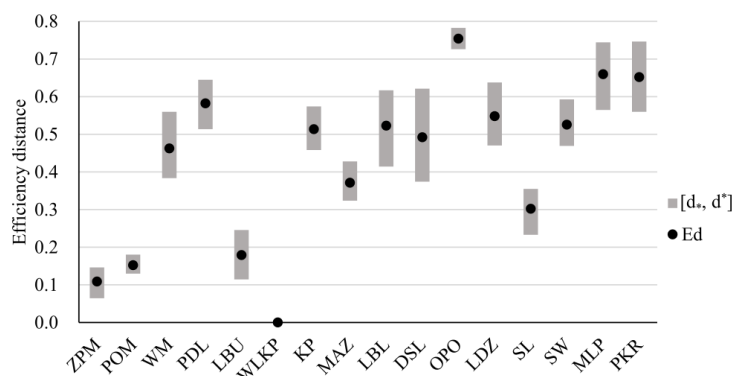


Figure 9. Extreme and expected distances to the best unit for Polish voivodeships when considering inhabitants' health improvement level.

Table 6 presents the distribution of distances for all voivodeships when considering the inhabitants' health improvement level. The only two provinces for which this distance is lower than 0.1 are WLKP (100%) and ZPM (33.2%). Such a favorable result was not attained by ZPM at the comprehensive level for any feasible weight vector. Then, its distance could drop even above 0.4. For the two units mentioned above, as well as POM and LBU, all samples confirm distances not higher than 0.2. Furthermore, for all provinces, we can indicate a single bucket in which the unit's distance falls for most samples. For example, it is (0.2, 0.3] for LBU, (0.3, 0.4] for MAZ, and (0.7, 0.8] for OPO. For these three provinces, the predominating distance buckets at the comprehensive level were better. However, for other voivodeships, including SW, MLP, and PKR, the most often repeated distance range worsened when limiting the analysis to the inhabitants' health improvement level.

The extreme and expected efficiency ranks at the inhabitants' health improvement level are provided in Figure 10. WLKP, as the only efficient province, is always ranked first. Both ZPM and POM are ranked between second and fourth, while for LBU, this range is slightly wider ([2, 5]). Conversely, only two voivodeships—OPO and PKR—fall to the bottom ranking at any point, and three others (WM, DSL, and MLP) are ranked fifteenth in the least favorable scenario. Finally, for DSL and MAZ, the difference between their extreme ranks is the greatest. For example, DSL is ranked between sixth and fifteenth. According to the expected ranks, WLKP ($ER = 1$), ZPM (2.110), POM (3.034), and LBU (3.860) are the best, and PKR (14.205), MLP (14.339), and OPO (15.990) are the worst.

Table 6. Distribution of distances to the best unit for Polish voivodeships when considering inhabitants’ health improvement level.

<i>DAII</i>	[0.0, 0.1]	(0.1, 0.2]	(0.2, 0.3]	(0.3, 0.4]	(0.4, 0.5]	(0.5, 0.6]	(0.6, 0.7]	(0.7, 0.8]	(0.8, 0.9]	(0.9, 1.0]
ZPM	0.332	0.668	0	0	0	0	0	0	0	0
POM	0	1	0	0	0	0	0	0	0	0
WM	0	0	0	0.055	0.727	0.218	0	0	0	0
PDL	0	0	0	0	0	0.658	0.342	0	0	0
LBU	0	0.713	0.287	0	0	0	0	0	0	0
WLKP	1	0	0	0	0	0	0	0	0	0
KP	0	0	0	0	0.341	0.659	0	0	0	0
MAZ	0	0	0	0.817	0.183	0	0	0	0	0
LBL	0	0	0	0	0.342	0.618	0.040	0	0	0
DSL	0	0	0	0.033	0.510	0.433	0.024	0	0	0
OPO	0	0	0	0	0	0	0	1	0	0
LDZ	0	0	0	0	0.150	0.710	0.140	0	0	0
SL	0	0	0.433	0.567	0	0	0	0	0	0
SW	0	0	0	0	0.256	0.744	0	0	0	0
MLP	0	0	0	0	0	0.088	0.715	0.197	0	0
PKR	0	0	0	0	0	0.125	0.708	0.167	0	0

When compared to the comprehensive level, the greatest improvement in the attained ranks can be observed for ZPM ([2, 4] rather than [4, 15]) and SL ([4, 5] rather than [7, 16]). On the contrary, the greatest deterioration of possible positions is noted for OPO ([15, 16] rather than [8, 16]) and MLP ([11, 16] rather than [4, 15]). For many provinces, including WM, POL, MAZ, LBL, LDZ, and SW, the ranking intervals got significantly narrower, confirming the lower diversity of results when limiting the scope of the analysis to health improvement.

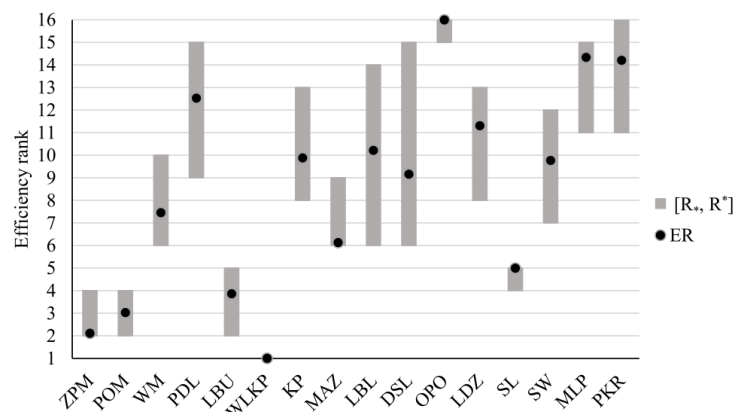


Figure 10. Extreme and expected ranks for Polish voivodeships when considering inhabitants’ health improvement level.

The distribution of efficiency ranks given the inhabitants’ health improvement level is presented in Table 7. The most stable individual positions were observed for WLKP (1–100%), SL (5–99.6%), OPO (16–99.0%), and ZPM (2–91.6%). Such high acceptabilities were not observed at the comprehensive level for any position and unit. In fact, the greatest share of weights (72.4%) supported the third position of LBU. Returning to the health improvement category, for some other voivodeships, the vast majority of weights indicate a pair of ranks (e.g., MLP and PKR are ranked 14th or 15th for 92.3% or 78.3% samples, respectively). Finally, the ranks of some units are more dependent on the chosen weight vector. For example, KP attained positions between 8th and 12th, with no *ERAI* exceeding 28%. Similarly, SW is ranked within the range [7, 12] with *ERAI*s not less than

6.9% and not greater than 31.1%. Still, these outcomes exhibit less diversity than at the comprehensive level.

Table 7. Distribution of ranks for Polish voivodeships when considering inhabitants’ health improvement level.

ERAI	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
ZPM	0	0.916	0.058	0.026	0	0	0	0	0	0	0	0	0	0	0	0
POM	0	0.054	0.858	0.088	0	0	0	0	0	0	0	0	0	0	0	0
WM	0	0	0	0	0	0.029	0.705	0.110	0.093	0.063	0	0	0	0	0	0
PDL	0	0	0	0	0	0	0	0	0.003	0.179	0.096	0.075	0.403	0.142	0.102	0
LBU	0	0.030	0.084	0.882	0.004	0	0	0	0	0	0	0	0	0	0	0
WLKP	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KP	0	0	0	0	0	0	0	0.212	0.272	0.182	0.087	0.247	0	0	0	0
MAZ	0	0	0	0	0	0.932	0.013	0.046	0.009	0	0	0	0	0	0	0
LBL	0	0	0	0	0	0	0.049	0.036	0.294	0.206	0.177	0.176	0.049	0.013	0	0
DSL	0	0	0	0	0	0.039	0.039	0.399	0.201	0.070	0.142	0.064	0.019	0.026	0.001	0
OPO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.010	0.990
LDZ	0	0	0	0	0	0	0	0.120	0.059	0.085	0.159	0.281	0.296	0	0	0
SL	0	0	0	0.004	0.996	0	0	0	0	0	0	0	0	0	0	0
SW	0	0	0	0	0	0	0.194	0.077	0.069	0.215	0.311	0.134	0	0	0	0
MLP	0	0	0	0	0	0	0	0	0	0	0.023	0.011	0.043	0.450	0.473	0
PKR	0	0	0	0	0	0	0	0	0	0	0.005	0.012	0.190	0.369	0.414	0.010

The graph of the necessary relation at the inhabitants’ health improvement level is shown in Figure 11. The robust conclusions are richer than at the comprehensive level. For example, the number of pairs of different provinces that are related by the necessary preference increased from 47 to 85. Moreover, the number of levels in the respective Hasse diagram increased from 3 to 7.

In particular, WLKP is necessarily preferred to all other provinces, and four other voivodeships—POM, ZPM, LBU, and SL—proves to be necessarily better than the remaining eleven units. The worst unit is OPO, which is not necessarily preferred to any other province while being possibly preferred only to PKR. Further, MLP and PKR are necessarily worse than 10 and 9 other provinces, respectively. Finally, SL benefited the most from limiting the scope to the health improvement level, because in the hierarchy’s root, it was not robustly better than any other voivodeship.

The respective *PEOIs* are given in Table 8. Similarly, as at the comprehensive level, for some pairs, one unit is significantly better than the other (see, e.g., ZPM and POM with $PEOI(ZPM, POM) = 94.8%$ or LBU and SL with $PEOI^H(LBU, SL) = 98.7%$). In turn, other pairs are characterized by more balanced stochastic acceptabilities, indicating an advantage of either voivodeship (see, e.g., PKR and MLP with $PEOI^H(PKR, MLP) = 52.2%$ or KP and SW with $PEOI^H(KP, SW) = 52.4%$). However, the absolute values of *PEOIs* for some pairs differ vastly compared to the hierarchy’s root. For example, POM is necessarily better than ZPM at the comprehensive level, but when considering only health improvement, ZPM attains no worse efficiency for 94.8% of feasible weights.

5.3. The Category of Financial Management

In this section, we discuss the results attained at the level of financial management. Instead of comparing them to the outcomes at the comprehensive level, we emphasize how managers can use them to improve the relative efficiency of provinces. Similar improvement strategies can be designed for other categories or hierarchy nodes.

Figure 12 presents the extreme and expected distances to the best province when limiting the scope to financial management. The most important result derived from their analysis is the division of provinces into efficient and inefficient. The minimal distance equals zero only for two units: POM and LBU. This means that they are the best performers among the sixteen voivodeships for at least one feasible weight vector. However, LBU

has slightly better maximal and expected distances to the best province than POM. When looking at their inputs and outputs within the financial category, they perform equally well (51.3) on output F_3 . It is the best value among all provinces. Moreover, the other two financial factors are greater for LBU, confirming that it transforms more beds (input F_2) into a more significant profit (output F_1). These two voivodeships should serve as ultimate peers for the remaining inefficient units in terms of financial management.

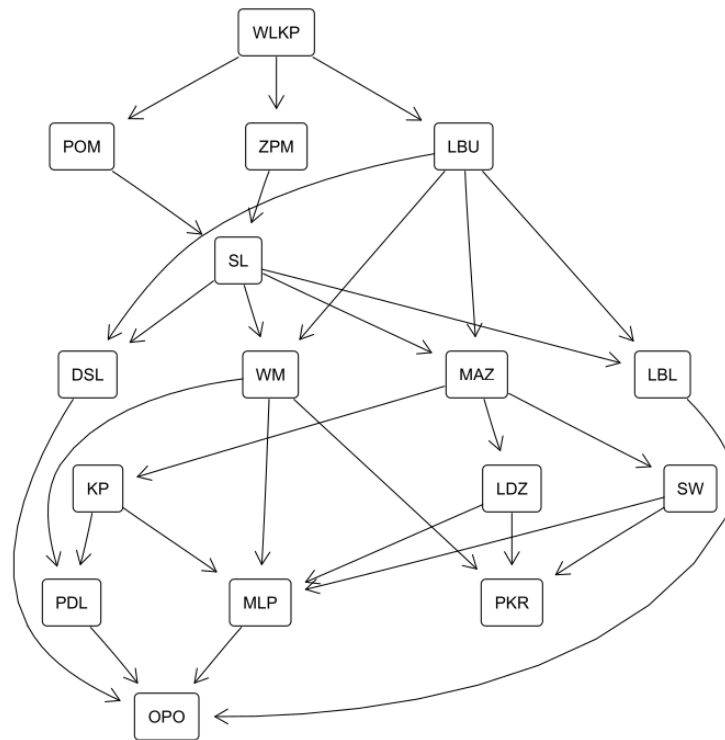


Figure 11. The Hasse diagram of the necessary efficiency preference relation when considering inhabitants’ health improvement level.

Table 8. Pairwise efficiency outranking indices for pairs of Polish voivodeships when considering inhabitants’ health improvement level.

PEOI	ZPM	POM	WM	PDL	LBU	WLKP	KP	MAZ	LBL	DSL	OPO	LDZ	SL	SW	MLP	PKR
ZPM	1	0.948	1	1	0.973	0	1	1	1	1	1	1	1	1	1	1
POM	0.052	1	1	1	0.897	0	1	1	1	1	1	1	1	1	1	1
WM	0	0	1	1	0	0	0.995	0.070	0.904	0.921	1	0.851	0	0.786	1	1
PDL	0	0	0	1	0	0	0	0	0.240	0.014	1	0.320	0	0.209	0.844	0.767
LBU	0.027	0.103	1	1	1	0	1	1	1	1	1	1	0.987	1	1	1
WLKP	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
KP	0	0	0.005	1	0	0	1	0	0.565	0.395	1	0.640	0	0.524	1	0.996
MAZ	0	0	0.930	1	0	0	1	1	0.989	0.939	1	1	0	1	1	1
LBL	0	0	0.096	0.760	0	0	0.435	0.011	1	0.278	1	0.688	0	0.489	0.962	1
DSL	0	0	0.079	0.986	0	0	0.605	0.061	0.722	1	1	0.728	0	0.635	0.942	0.947
OPO	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0.015
LDZ	0	0	0.149	0.680	0	0	0.360	0	0.312	0.272	1	1	0	0.011	1	1
SL	0	0	1	1	0.013	0	1	1	1	1	1	1	1	1	1	1
SW	0	0	0.214	0.791	0	0	0.476	0	0.511	0.365	1	0.989	0	1	1	1
MLP	0	0	0	0.156	0	0	0	0	0.038	0.058	1	0	0	0	1	0.478
PKR	0	0	0	0.233	0	0	0.004	0	0	0.053	0.985	0	0	0	0.522	1

Among them, the overall good performers are WLKP ($Ed^F = 0.1689$) and MLP ($Ed^F = 0.2328$). They do not optimize one specific input or output, but perform decently on all indicators. In the optimistic, pessimistic, and expected scenarios, SL attains the least favorable results with the significantly greatest distances to the best province. In turn, KP

has the broadest range of distances ([0.140, 0.416]), confirming its performance’s sensitivity to the selection of particular priorities. This results from an imbalanced performance profile with a highly favorable value on output F_1 and a relatively poor value on output F_3 . Finally, when complete order is desired, it can be imposed by the expected distances. In this case, LBU ($Ed^F = 0.010$) and POM ($Ed^F = 0.021$) are safely ranked at the top, and DSL ($Ed^F = 0.509$) and SL ($Ed^F = 0.735$) are ranked at the bottom.

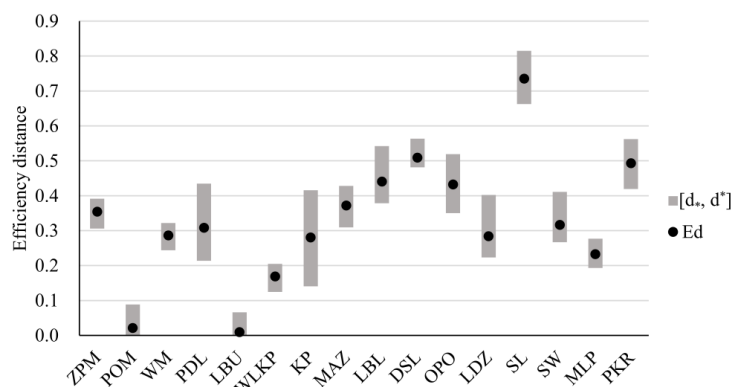


Figure 12. Extreme and expected distances to the best unit for Polish voivodeships when considering financial management level.

The distance distribution at the financial management level is presented in Table 9. For both efficient provinces (POM and LBU), the distance is always within the first bucket ([0.0, 0.1]). Among the inefficient provinces, WLKP confirms its superiority over the remaining units, as for over 99% samples, its distance from the efficient unit is not greater than 0.2. The only two other provinces with positive $DAIIs$ for bucket (0.1, 0.2] are KP (14.2%) and MLP (4.3%). The greatest stability of distances among inefficient provinces is observed for ZPM with $DAII^F(ZPM, (0.3, 0.4]) = 1$. There are only two provinces, PDL and KP, for which most samples confirm no single distance bucket. Furthermore, KP is the only unit with positive $DAIIs$ for more than three buckets ((0.1, 0.5]).

Table 9. Distribution of distances to the best unit for Polish voivodeships when considering financial management level.

<i>DAII</i>	[0.0, 0.1]	(0.1, 0.2]	(0.2, 0.3]	(0.3, 0.4]	(0.4, 0.5]	(0.5, 0.6]	(0.6, 0.7]	(0.7, 0.8]	(0.8, 0.9]	(0.9, 1.0]
ZPM	0	0	0	1	0	0	0	0	0	0
POM	1	0	0	0	0	0	0	0	0	0
WM	0	0	0.757	0.243	0	0	0	0	0	0
PDL	0	0	0.475	0.483	0.042	0	0	0	0	0
LBU	1	0	0	0	0	0	0	0	0	0
WLKP	0	0.993	0.007	0	0	0	0	0	0	0
KP	0	0.142	0.453	0.379	0.026	0	0	0	0	0
MAZ	0	0	0	0.857	0.143	0	0	0	0	0
LBL	0	0	0	0.139	0.775	0.086	0	0	0	0
DSL	0	0	0	0	0.312	0.688	0	0	0	0
OPO	0	0	0	0.240	0.708	0.052	0	0	0	0
LDZ	0	0	0.705	0.295	0	0	0	0	0	0
SL	0	0	0	0	0	0	0.220	0.743	0.037	0
SW	0	0	0.366	0.628	0.006	0	0	0	0	0
MLP	0	0.043	0.957	0	0	0	0	0	0	0
PKR	0	0	0	0	0.545	0.455	0	0	0	0

The results of the robustness analysis for efficiency ranks at the financial management level are presented in Figure 13 (extreme and expected positions) and Table 10 (efficiency

rank acceptability indices). They provide additional insights into the comparisons of efficient provinces. Specifically, even if LUB and POM can be ranked first for some weight vectors, the former is ranked first almost twice as often as the latter. This makes it the most favorable province regarding financial management, as additionally confirmed by the expected ranks ($ER^F(LUB) = 1.377$ vs. $ER^F(POM) = 1.623$). The remaining provinces are at most third in the best case. Again, WLKP proves to be the most advantageous among them, with $ER^F(WLKP) = 3.107$. Further, KP has the broadest range of efficiency positions, being ranked between third and eleventh. $ERAI$ s confirm that its ranks are rather equally distributed between these extreme positions, with the maximal value for the fifth rank ($ERAI^F(KP, 5) = 24.8\%$) and acceptabilities greater than 6% for all ranks within the range [3, 11]. Such great diversity is a consequence of its extreme performances. In fact, it is the best among all provinces on F_1 while being in the bottom three on F_3 . Thus, KP is focused on the profit attained by healthcare institutions rather than on the number of treated patients. This aspect needs to be improved when aiming for higher ranks.

There are three other provinces with relatively wide possible efficiency rank intervals: ZPM ([6, 12]), PDL ([5, 11]), and OPO ([9, 15]). Among them, only ZPM attains a single rank for most samples ($ERAI^F(ZPM, 10) = 62.4\%$). Finally, SL is ranked at the bottom regardless of the weight vector. This is related to its greatest value on input F_2 , the lowest value on output F_3 , and a relatively low value on output F_1 . Thus, even if the financial input of SL is the greatest, its outputs are less favorable than for provinces with lesser financial resources. The complete ranking established with the expected efficiency ranks aligns with the one based on expected distances, with LBU, POM, WLKP, and MLP being ranked among the best provinces and SL, DSL, and PKR placed at the bottom.

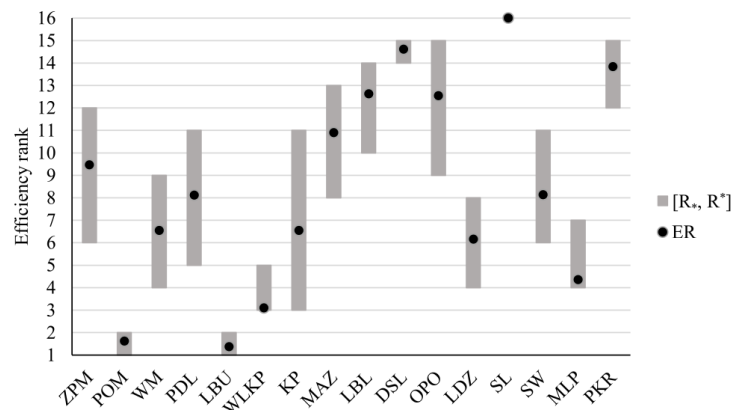


Figure 13. Extreme and expected ranks for Polish voivodeships when considering financial management level.

Figure 14 presents the necessary ranking at the financial management level. POM and LBU are necessarily preferred to all other provinces, confirming their superiority given the financial category. The second level includes KP, WLKP, and MLP, and the lowest one contains PKR, DSL, and SL. They are possibly preferred to 4, 3, and 0 other units, respectively. Moreover, SL is necessarily worse than all other voivodeships.

The pairwise comparisons are useful for analysts particularly familiar with some provinces. For example, the authorities of OPO can compare it to other provinces searching for possible improvements. They can note that OPO is robustly worse than PDL or LBU and robustly better than SL. Notably, the necessary ranking (see Figure 14) is a good starting point to find the improvement paths for provinces. OPO has multiple paths to achieve efficiency. For example, it can take PDL as the first benchmark, follow WLKP, and finally refer to POM or LBU. An alternative improvement path runs through KP and POM.

Table 10. Distribution of ranks for Polish voivodeships when considering financial management level.

ERAI	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
ZPM	0	0	0	0	0	0.011	0.046	0.182	0.062	0.624	0.068	0.007	0	0	0	0
POM	0.377	0.623	0	0	0	0	0	0	0	0	0	0	0	0	0	0
WM	0	0	0	0.048	0.318	0.207	0.089	0.143	0.195	0	0	0	0	0	0	0
PDL	0	0	0	0	0.001	0.206	0.200	0.205	0.133	0.164	0.091	0	0	0	0	0
LBU	0.623	0.377	0	0	0	0	0	0	0	0	0	0	0	0	0	0
WLKP	0	0	0.900	0.093	0.007	0	0	0	0	0	0	0	0	0	0	0
KP	0	0	0.100	0.088	0.248	0.169	0.066	0.077	0.074	0.061	0.117	0	0	0	0	0
MAZ	0	0	0	0	0	0	0	0.003	0.151	0.054	0.563	0.187	0.042	0	0	0
LBL	0	0	0	0	0	0	0	0	0	0.012	0.046	0.379	0.424	0.139	0	0
DSL	0	0	0	0	0	0	0	0	0	0	0	0	0	0.382	0.618	0
OPO	0	0	0	0	0	0	0	0	0.005	0.059	0.114	0.300	0.339	0.092	0.091	0
LDZ	0	0	0	0.100	0.126	0.365	0.332	0.077	0	0	0	0	0	0	0	0
SL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
SW	0	0	0	0	0	0.013	0.267	0.313	0.380	0.026	0.001	0	0	0	0	0
MLP	0	0	0	0.671	0.300	0.029	0	0	0	0	0	0	0	0	0	0
PKR	0	0	0	0	0	0	0	0	0	0	0	0.127	0.195	0.387	0.291	0

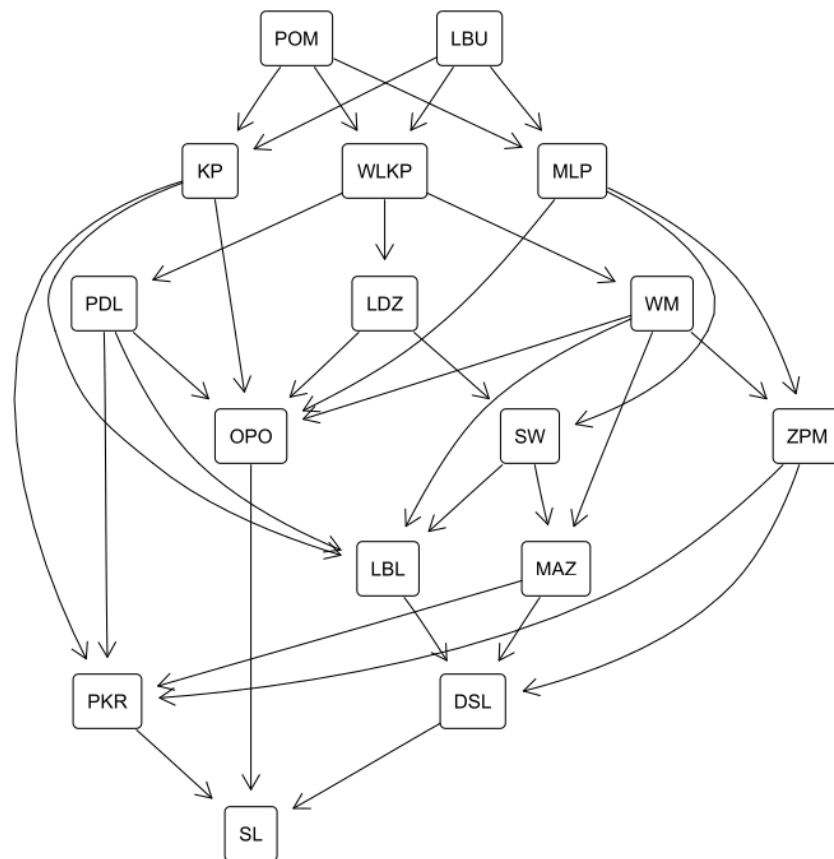


Figure 14. The Hasse diagram of the necessary efficiency preference relation when considering financial management level.

The analysis of $PEOIs$ (see Table 11) is helpful for pairs related by mutual possible preference. For some, one province attains greater efficiency more often (e.g., $PEOI^F(SW, ZPM) = 91.2\%$). For other pairs, indicating a better voivodeship is more challenging, as the shares of weights confirming the advantage of either unit are similar. An example of such a pair is LDZ and KP with $PEOI^F(LDZ, KP) = 48.3\%$ and $PEOI^F(KP, LDZ) = 51.7\%$.

Table 11. Pairwise efficiency outranking indices for pairs of Polish voivodeships when considering financial management level.

PEOI	ZPM	POM	WM	PDL	LBU	WLKP	KP	MAZ	LBL	DSL	OPO	LDZ	SL	SW	MLP	PKR
ZPM	1	0	0	0.285	0	0	0.244	0.999	0.986	1	0.925	0.012	1	0.088	0	1
POM	1	1	1	1	0.403	1	1	1	1	1	1	1	1	1	1	1
WM	1	0	1	0.661	0	0	0.512	1	1	1	1	0.510	1	0.816	0.083	1
PDL	0.715	0	0.339	1	0	0	0.197	0.800	1	1	1	0.251	1	0.546	0.005	1
LBU	1	0.597	1	1	1	1	1	1	1	1	1	1	1	1	1	1
WLKP	1	0	1	1	0	1	0.920	1	1	1	1	1	1	1	0.996	1
KP	0.756	0	0.488	0.803	0	0.080	1	0.813	1	1	1	0.517	1	0.665	0.172	1
MAZ	0.001	0	0	0.200	0	0	0.187	1	0.932	1	0.809	0	1	0	0	1
LBL	0.014	0	0	0	0	0	0	0.068	1	1	0.480	0	1	0	0	0.752
DSL	0	0	0	0	0	0	0	0	0	1	0.102	0	1	0	0	0.287
OPO	0.075	0	0	0	0	0	0	0.191	0.520	0.898	1	0	1	0.006	0	0.754
LDZ	0.988	0	0.490	0.749	0	0	0.483	1	1	1	1	1	1	1	0.109	1
SL	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
SW	0.912	0	0.184	0.454	0	0	0.335	1	1	1	0.994	0	1	1	0	1
MLP	1	0	0.917	0.995	0	0.004	0.828	1	1	1	1	0.891	1	1	1	1
PKR	0	0	0	0	0	0	0	0	0.248	0.713	0.246	0	1	0	0	1

5.4. The Category of Consumer Satisfaction

In this section, we present the results of provinces' performance analysis at the level of consumer satisfaction with the healthcare system. To save space, we refer only to the exact and expected results without reporting the complete tables with stochastic acceptabilities.

Figure 15 presents the extreme and expected distances when considering the satisfaction of consumers. There are five efficient provinces with $d_*^S = 0$: POM, KP, LBL, SW, and PKR. This is the greatest number for any node in the considered hierarchy. Among them, two provinces, LBL and PKR, have particularly narrow distance intervals, being close to the efficient units regardless of the weight vector ($d^{S,*} = 0.081$ for LBL and $d^{S,*} = 0.057$ for PKR). They are also the best regarding the expected distance (for PKR— $Ed^S = 0.005$ and for LBL— $Ed^S = 0.029$). In turn, the maximal (worst) possible distances for SW (0.285) and POM (0.307) are vastly greater, emphasizing their sensitivity to the selection of a particular weight vector. Among the inefficient units, the best minimal (optimistic) distance is achieved by OPO (0.049), and the best maximal (pessimistic) distance is attained by WLKP (0.312). The three provinces that are the worst considering both minimal and maximal distance are PDL ($d_*^S = 0.454$, $d^{S,*} = 0.556$), DSL ($d_*^S = 0.449$, $d^{S,*} = 0.623$), and ZPM ($d_*^S = 0.513$, $d^{S,*} = 0.670$). Their poor performance is also reflected in the bottom ranks according to the expected distances.

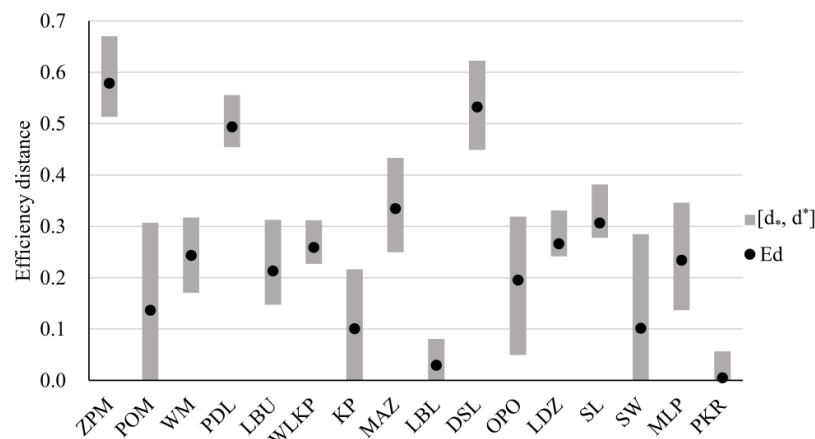


Figure 15. Extreme and expected distances to the best unit for Polish voivodeships when considering consumer satisfaction level.

The extreme and expected efficiency ranks in consumer satisfaction are presented in Figure 16. Among the five efficient provinces, PKR has the narrowest rank interval; in the least favorable scenario, it is ranked third. On the contrary, even if POM is efficient, it can drop even to the tenth position, making its performance the least stable among all voivodeships. LBU is the best inefficient province when it comes to the best rank ($R_*^S = 3$), and the best-ranked inefficient units in the pessimistic settings are WM, LBU, and OPO ($R^{S,*} = 9$). The worst provinces considering both the minimal and maximal ranks are ZPM ([15, 16]), PDL ([14, 15]), and DSL ([14, 16]). Notably, their rank intervals are very narrow. As far as the expected ranks are concerned, the best province is PKR (1.426), followed by LBL (2.249) and KP (3.441). On the other extreme, PDL (14.192), DSL (14.988), and ZPM (15.820) are the least favorable.

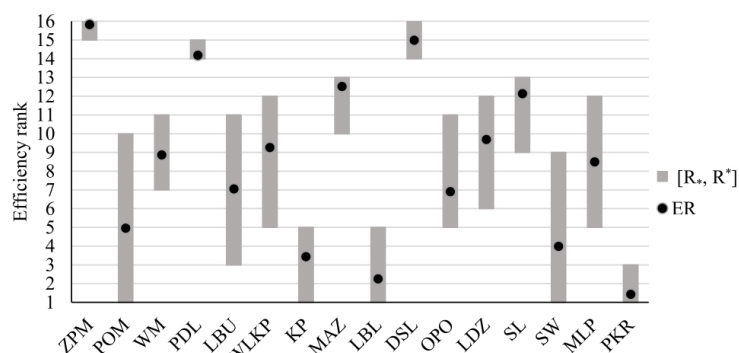


Figure 16. Extreme and expected ranks for Polish voivodeships when considering consumer satisfaction level.

The necessary efficiency preference relation is presented as the Hasse diagram in Figure 17. The provinces form the multiple-level structure, with efficient ones (POM, PKR, LBL, KP, and SW) at the top. Among them, PKR and LBL are necessarily preferred to the most significant number of eleven other provinces, while POM proves robustly better than only six other provinces. Five other voivodeships (WM, LBU, OPO, WLKP, and MLP) are placed in the second level. All of them are necessarily preferred to four or five other provinces. Finally, DSL and ZPM are confirmed as the worst provinces, as they are not necessarily preferred to any other province. In fact, DSL is possibly preferred to only two other units (ZPM and PDL), while ZPM is at least as efficient for at least one weight vector only when compared to DSL.

5.5. Complete Efficiency Rankings of Voivodeships

The expected distances (*Ed*) and ranks (*ER*) allow for the construction of a complete ranking of voivodeships. In this section, we compare such orders with the ones obtained with the most commonly used ranking procedures for DEA, i.e., Cross-efficiency (CE) [30] and Super-efficiency (SE) [31]. We adapted them to a value-based additive efficiency model and ran it on each category of indicators.

The rankings generated by all four procedures for the level of inhabitants' health improvement are provided in Table 12. To quantify the agreement level of such rankings for all hierarchy nodes, we used Kendall's τ coefficient [28,32]. Its values are shown in Table 13. Note that -1 means the rankings are inverse, whereas 1 denotes a pair of the same rankings.

All four procedures provide highly correlated rankings. The two methods proposed in this paper (*Ed* and *ER*) offer the most similar orders of provinces. The rankings constructed with these two procedures are the same for the comprehensive analysis of the healthcare systems and the customer satisfaction category. For the remaining two categories, Kendall's τ coefficient equals 0.97.

The similarity between the rankings based on *Ed*, *ER*, and CE is between 0.73 for the comprehensive analysis and 0.97 for health improvement and financial management

perspectives. The average measure value is 0.904. When comparing the orders imposed by SE and the two robustness-based methods, Kendall's τ is between 0.72 and 0.95 (0.92) for Ed (ER), with an average value of 0.879 (0.863). Given that the maximal possible value of Kendall's τ is 1, the observed similarity levels are very high.

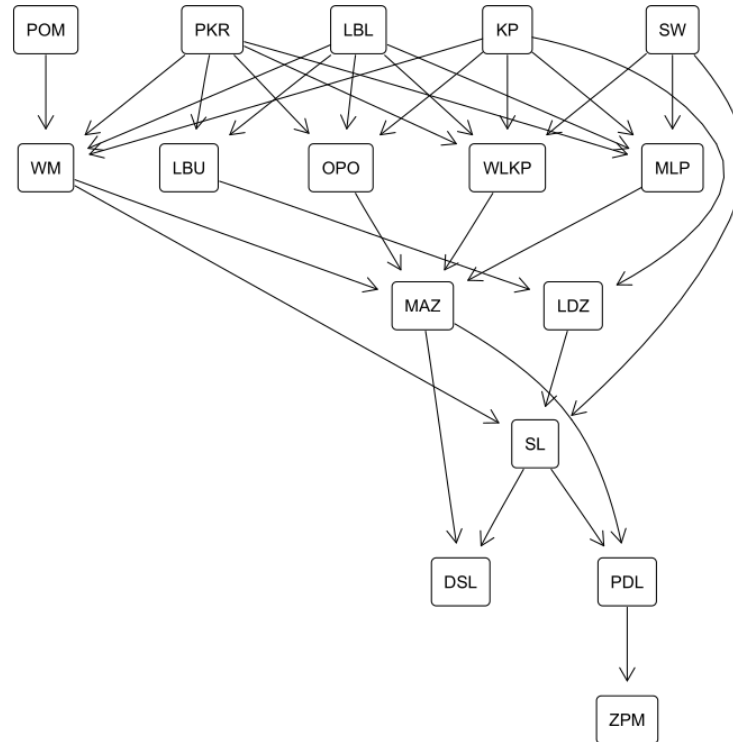


Figure 17. The Hasse diagram of the necessary efficiency preference relation when considering consumer satisfaction level.

Table 12. The voivodeships' efficiency rankings imposed by four different measures at the level of health improvement: expected distance (Ed), expected rank (ER), cross-efficiency (CE), and super-efficiency (SE).

	ZPM	POM	WM	PDL	LBU	WLKP	KP	MAZ	LBL	DSL	OPO	LDZ	SL	SW	MLP	PKR
Ed	0.109	0.152	0.462	0.582	0.179	0.000	0.514	0.372	0.523	0.492	0.754	0.548	0.302	0.526	0.659	0.652
	2	3	7	13	4	1	9	6	10	8	16	12	5	11	15	14
ER	2.110	3.034	7.456	12.530	3.860	1.000	9.885	6.132	10.215	9.164	15.990	11.310	4.996	9.774	14.339	14.205
	2	3	7	13	4	1	10	6	11	8	16	12	5	9	15	14
CE	0.102	0.155	0.471	0.598	0.178	0.000	0.525	0.378	0.507	0.495	0.760	0.526	0.294	0.509	0.656	0.630
	2	3	7	13	4	1	11	6	9	8	16	12	5	10	15	14
SE	0.065	0.130	0.383	0.514	0.114	-0.140	0.458	0.324	0.414	0.374	0.726	0.470	0.233	0.469	0.565	0.560
	2	4	8	13	3	1	10	6	9	7	16	12	5	11	15	14

Table 13. Kendall's τ coefficient for ranking procedures at each hierarchy level.

	Comprehensive Analysis				Inhabitants' Health Improvement				Effective Financial Management				Consumer Satisfaction			
	Ed	ER	CE	SE	Ed	ER	CE	SE	Ed	ER	CE	SE	Ed	ER	CE	SE
Ed	1.00	1.00	0.73	0.72	1.00	0.97	0.97	0.95	1.00	0.97	0.97	0.95	1.00	1.00	0.95	0.90
ER	1.00	1.00	0.73	0.72	0.97	1.00	0.97	0.92	0.97	1.00	0.97	0.92	1.00	1.00	0.95	0.90
CE	0.73	0.73	1.00	0.85	0.97	0.97	1.00	0.95	0.97	0.97	1.00	0.92	0.95	0.95	1.00	0.92
SE	0.72	0.72	0.85	1.00	0.95	0.92	0.95	1.00	0.95	0.92	0.92	1.00	0.90	0.90	0.92	1.00

5.6. Discussion

Three provinces, namely POM, WLKP, and LBU, prove to be the best in the comprehensive analysis of the healthcare system in Poland. However, the results for the three subcategories differ. This section discusses the conclusions that can be derived from the cross-category analysis, indicating various provinces' strong and weak points. For illustrative purposes, we refer to three example voivodeships representing the top (WLKP), middle (PKR), and bottom (OPO) performers in the hierarchy's root.

When it comes to WLKP, it proves to be the best at the inhabitants' health improvement level; it is necessarily preferred by POM and LBU for the financial management level, and its ranks are between fifth and twelfth, given the consumer satisfaction perspective. This suggests that despite the decent quality of medical services, there is room for improvement in patient satisfaction and financial management. In particular, managers can conduct some training in soft skills for medical staff to improve consumer assessment of the healthcare system.

PKR is the best province regarding consumer satisfaction, as confirmed by its favorable expected distance and rank. However, when considering the comprehensive results and conclusions drawn for the remaining two categories, it performs poorly. Its expected rank is greater than 13 for inhabitants' health improvement and financial management categories, while in the hierarchy's root, its average rank is 12.024. Hence, the healthcare system managers in this province should focus on improvements in medical decisions and financial management. In turn, other provinces can consider the healthcare system in PKR as the benchmark of proper communication with the consumers.

Such a cross-category analysis can serve as the basis for designing an improvement plan for provinces that proved to be relatively bad in the comprehensive analysis. In particular, the analysis of OPO's poor performance points to the inhabitants' health improvement category and financial management. It is the worst province for the former perspective for 99% of weight vectors, and its expected rank is only 12.549 for the latter category. Hence, managers should first focus on improving inhabitants' health, due to the high importance of this category in the analysis. Then, they should design a plan for advancing financial management.

6. Conclusions

This paper introduces a novel framework for robustness analysis in the context of additive value-based efficiency analysis with a hierarchical structure. It admits a multiple-layer organization of relevant factors from the most general to the most detailed ones while tolerating both inputs and outputs in the same node. We accept the linear weight restrictions concerning the importance and trade-offs between various factors or subcategories with the common predecessor in the hierarchy. The results can be considered in each hierarchy node, letting the analyst view the comprehensive outcomes and draw conclusions for the subproblems where the relevant factors are limited to concise subsets of inputs and outputs reflecting a particular perspective. The proposed framework can be used in the standard efficiency analysis and the decision contexts requiring the consideration of composite indicators.

We derived the results by considering three perspectives: score-based distances to the efficient unit, ranks, and pairwise preference relations. For each of them, we proposed a pair of methods. The first group was based on mathematical programming, offering the exact, extreme outcomes that can be attained in the set of feasible input/output weights. The other group was based on Monte Carlo simulations, providing the distribution of efficiency outcomes through stochastic acceptabilities that estimate the share of the weight subspaces confirming a given result. These approaches are complementary because the exact outcomes often need more conclusiveness, whereas the stochastic indices—even if approximated with high accuracy—may fail to capture some extreme results.

We illustrated the framework's applicability by assessing the quality of the healthcare system of sixteen Polish voivodeships. The analysis included nine indicators of different

natures and concerned four levels. The comprehensive results were based on all relevant characteristics considered jointly, while the three subproblems captured the perspectives of inhabitants' health improvement, financial management, and consumer satisfaction. We reported three provinces—Pomorskie (POM), Wielkopolskie (WLKP), and Lubuskie (LBU)—as the most efficient ones. We presented their strong and weak points by referring to the results in all hierarchy nodes. Moreover, we discussed the practical usefulness of the robustness analysis in terms of managerial implications.

Author Contributions: A.L.-K.: conceptualization, methodology, software, investigation, data curation, writing—original draft, visualization; M.K.: conceptualization, methodology, validation, formal analysis, investigation, writing—original draft, supervision, project administration, funding acquisition; W.M.: conceptualization, software, visualization, writing—revision. All authors have read and agreed to the published version of the manuscript.

Funding: Anna Labijak-Kowalska is grateful for the support from the Polish Ministry of Education and Science (grant no. 0311/SBAD/0735). Miłosz Kadziński acknowledges support from the Polish National Science Center under the SONATA BIS project (grant no. DEC-2019/34/E/HS4/00045).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The complete research data produced within the study is contained within this article.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AHP	Analytical Hierarchy Process
CE	Cross-efficiency
DEA	Data Envelopment Analysis
DMU	Decision Making Unit
LP	Linear Programming
MAVT	Multi-Attribute Value Theory
MCDA	Multiple Criteria Decision Analysis
MCHP	Multiple Criteria Hierarchy Process
MILP	Mixed-Integer Linear Programming
SE	Super-efficiency

Appendix A. Properties of the Exact Robust Results

This section presents the relevant properties of the exact robust results given the hierarchical structure. In particular, when the minimal distances in all children nodes of some more general category are equal to zero, the minimal distance in this category must also be zero.

Proposition A1. For $DMU_o \in \mathcal{D}$ and category $c_i^{(l)} \in \mathbf{N} \setminus \mathbf{f}$, if $\forall t \in ch(c_i^{(l)}) : d_{*,o}^t = 0$, then $d_{*,o}^{c_i^{(l)}} = 0$.

Similarly, if the minimal rank of DMU_o in all children nodes of some more general category is 1 (K), then it needs to be ranked first (last) in this category in the best case.

Proposition A2. For DMU_o and category $c_i^{(l)} \in \mathbf{N} \setminus \mathbf{f}$, if $\forall t \in ch(c_i^{(l)}) : R_{*,o}^t = 1$ then $R_{*,o}^{c_i^{(l)}} = 1$.

Proposition A3. For DMU_o and category $c_i^{(l)}$ in $N \setminus f$, if $\forall t \in ch(c_i^{(l)}) : R_{*,o}^t = K$, then $R_{*,o}^{c_i^{(l)}} = K$.

Moreover, if the maximal rank of DMU_o in all children nodes of some more general category is 1, then it needs to be ranked first in this category, even in the worst case.

Proposition A4. For DMU_o and category $c_i^{(l)} \in N \setminus f$, if $\forall t \in ch(c_i^{(l)}) : R_o^{*,t} = 1$, then $R_o^{*,c_i^{(l)}} = 1$.

When DMU_o is necessarily preferred to DMU_k in all children nodes of some more general category, then DMU_o must be necessarily preferred to DMU_k , given this category.

Proposition A5. For pair (DMU_o, DMU_k) and category $c_i^{(l)} \in N \setminus f$, if $\forall t \in ch(c_i^{(l)}) : DMU_o \succ_E^{N,t} DMU_k$, then $DMU_o \succ_E^{N,c_i^{(l)}} DMU_k$.

Note that $\succ_E^{N,c_i^{(l)}}$ is a partial preorder (i.e., transitive and reflexive). When DMU_o is not possibly preferred to DMU_k in all children nodes of some more general category, then DMU_o is not possibly preferred to DMU_k , given this category.

Proposition A6. For pair (DMU_o, DMU_k) and category $c_i^{(l)} \in N \setminus f$, if $\forall t \in ch(c_i^{(l)}) : \neg(DMU_o \succ_E^{P,t} DMU_k)$, then $\neg(DMU_o \succ_E^{P,c_i^{(l)}} DMU_k)$.

Moreover, when DMU_o is necessarily preferred to DMU_k in all children nodes of some more general category except one node for which it is possibly preferred, then DMU_o needs to be possibly preferred to DMU_k , given this category.

Proposition A7. For pair (DMU_o, DMU_k) and category $c_i^{(l)} \in N \setminus f$, if $\forall t \in ch(c_i^{(l)}) \setminus a : DMU_o \succ_E^{N,t} DMU_k \wedge DMU_o \succ_E^{P,a} DMU_k$, then $DMU_o \succ_E^{P,c_i^{(l)}} DMU_k$.

Note that $\succ_E^{P,c_i^{(l)}}$ is negatively transitive and strongly complete. Moreover, the truth of the necessary preference implies the truth of the possible preference ($\succ_E^{N,c_i^{(l)}} \subseteq \succ_E^{P,c_i^{(l)}}$).

Appendix B. Formulations of Example Mathematical Programming Models for Computing the Exact Robust Results

This section illustrates mathematical programming models for computing the exact robust results. They support understanding the general formulations presented in Section 4. We consider a hierarchical structure involving four indicators as presented in Figure 3. The input and output values of four DMUs and the indicator bounds are shown in Table A1. For simplicity, we assume that the marginal value functions are linear to make them easily computable. For example, for DMU_1 and input i_1 , the marginal value is calculated as $u_{i_1}(DMU_1) = u_{i_1}(9) = \frac{10-9}{10-0} = 0.1$. Analogously, for DMU_3 and output o_2 , the value of marginal function is equal to $u_{o_2}(DMU_3) = u_{o_2}(6) = \frac{6-5}{10-5} = 0.2$. Moreover, the following weight constraints are considered: $w_{C_1(1)} \geq w_{C_1(1)}, w_{C_2(1)} \geq 0.1, w_{i_1} \geq 0.2$, and $w_{o_2} \geq 0.6$.

Table A1. Input/output values for DMUs and indicators' bounds in the illustrative problem.

DMU	i_1	o_1	i_2	o_2
DMU ₁	9	5	8	7
DMU ₂	2	8	5	5
DMU ₃	5	6	7	6
DMU ₄	1	2	9	8
min	0	0	5	5
max	10	10	10	10

All formulations concern category $C_1^{(2)}$ (i.e., the hierarchy's root). Let us first provide the model for computing the minimal distance of DMU_1 to the best unit:

Minimize d_1

s.t.

$$\left. \begin{aligned}
 & \hat{w}_{i_1}^{C_1^{(2)}} u_{i_1}(9) + \hat{w}_{o_1}^{C_1^{(2)}} u_{o_1}(5) + \hat{w}_{i_2}^{C_1^{(2)}} u_{i_2}(8) + \hat{w}_{o_2}^{C_1^{(2)}} u_{o_2}(7) - (\hat{w}_{i_1}^{C_1^{(2)}} u_{i_1}(9) + \hat{w}_{o_1}^{C_1^{(2)}} u_{o_1}(5) + \hat{w}_{i_2}^{C_1^{(2)}} u_{i_2}(8) + \hat{w}_{o_2}^{C_1^{(2)}} u_{o_2}(7)) \leq d_1 \\
 & \hat{w}_{i_1}^{C_1^{(2)}} u_{i_1}(2) + \hat{w}_{o_1}^{C_1^{(2)}} u_{o_1}(8) + \hat{w}_{i_2}^{C_1^{(2)}} u_{i_2}(5) + \hat{w}_{o_2}^{C_1^{(2)}} u_{o_2}(5) - (\hat{w}_{i_1}^{C_1^{(2)}} u_{i_1}(9) + \hat{w}_{o_1}^{C_1^{(2)}} u_{o_1}(5) + \hat{w}_{i_2}^{C_1^{(2)}} u_{i_2}(8) + \hat{w}_{o_2}^{C_1^{(2)}} u_{o_2}(7)) \leq d_1 \\
 & \hat{w}_{i_1}^{C_1^{(2)}} u_{i_1}(5) + \hat{w}_{o_1}^{C_1^{(2)}} u_{o_1}(6) + \hat{w}_{i_2}^{C_1^{(2)}} u_{i_2}(7) + \hat{w}_{o_2}^{C_1^{(2)}} u_{o_2}(6) - (\hat{w}_{i_1}^{C_1^{(2)}} u_{i_1}(9) + \hat{w}_{o_1}^{C_1^{(2)}} u_{o_1}(5) + \hat{w}_{i_2}^{C_1^{(2)}} u_{i_2}(8) + \hat{w}_{o_2}^{C_1^{(2)}} u_{o_2}(7)) \leq d_1 \\
 & \hat{w}_{i_1}^{C_1^{(2)}} u_{i_1}(1) + \hat{w}_{o_1}^{C_1^{(2)}} u_{o_1}(2) + \hat{w}_{i_2}^{C_1^{(2)}} u_{i_2}(9) + \hat{w}_{o_2}^{C_1^{(2)}} u_{o_2}(8) - (\hat{w}_{i_1}^{C_1^{(2)}} u_{i_1}(9) + \hat{w}_{o_1}^{C_1^{(2)}} u_{o_1}(5) + \hat{w}_{i_2}^{C_1^{(2)}} u_{i_2}(8) + \hat{w}_{o_2}^{C_1^{(2)}} u_{o_2}(7)) \leq d_1 \\
 & d_1 \geq 0 \\
 & \hat{w}_{i_1}^{C_1^{(2)}} + \hat{w}_{o_1}^{C_1^{(2)}} + \hat{w}_{i_2}^{C_1^{(2)}} + \hat{w}_{o_2}^{C_1^{(2)}} = 1 \\
 & \frac{\hat{w}_{i_1}^{C_1^{(2)}} + \hat{w}_{o_1}^{C_1^{(2)}}}{\hat{w}_{i_1}^{C_1^{(2)}} + \hat{w}_{o_1}^{C_1^{(2)}} + \hat{w}_{i_2}^{C_1^{(2)}} + \hat{w}_{o_2}^{C_1^{(2)}}} \geq \frac{\hat{w}_{i_2}^{C_1^{(2)}} + \hat{w}_{o_2}^{C_1^{(2)}}}{\hat{w}_{i_1}^{C_1^{(2)}} + \hat{w}_{o_1}^{C_1^{(2)}} + \hat{w}_{i_2}^{C_1^{(2)}} + \hat{w}_{o_2}^{C_1^{(2)}}} \\
 & \frac{\hat{w}_{i_2}^{C_1^{(2)}} + \hat{w}_{o_2}^{C_1^{(2)}}}{\hat{w}_{i_1}^{C_1^{(2)}} + \hat{w}_{o_1}^{C_1^{(2)}} + \hat{w}_{i_2}^{C_1^{(2)}} + \hat{w}_{o_2}^{C_1^{(2)}}} \geq 0.1 \\
 & \frac{\hat{w}_{i_1}^{C_1^{(2)}}}{\hat{w}_{i_1}^{C_1^{(2)}} + \hat{w}_{o_1}^{C_1^{(2)}}} \geq 0.2 \\
 & \frac{\hat{w}_{o_2}^{C_1^{(2)}}}{\hat{w}_{i_1}^{C_1^{(2)}} + \hat{w}_{o_1}^{C_1^{(2)}}} \leq 0.6 \\
 & \hat{w}_{i_1}^{C_1^{(2)}} \geq 0, \hat{w}_{o_1}^{C_1^{(2)}} \geq 0, \hat{w}_{i_2}^{C_1^{(2)}} \geq 0, \hat{w}_{o_2}^{C_1^{(2)}} \geq 0.
 \end{aligned} \right\} \mathcal{W}^{C_1^{(2)}} \tag{A1}$$

In the above model, the first four constraints ensure that the optimized distance will correspond to the greatest distance of DMU_1 to other DMU. The fifth constraint ensures that the distance is non-negative. The constraint set $\mathcal{W}^{C_1^{(2)}}$ corresponds to the weight restrictions. In particular, the first constraint ensures that the sum of weights equals one, whereas the last guarantees that all weights are non-negative. In turn, to find the maximal distance of DMU_1 , we need to solve the following model:

Maximize d_1

s.t.

$$\left. \begin{aligned}
 & \hat{w}_{i_1}^{C_1^{(2)}} u_{i_1}(9) + \hat{w}_{o_1}^{C_1^{(2)}} u_{o_1}(5) + \hat{w}_{i_2}^{C_1^{(2)}} u_{i_2}(8) + \hat{w}_{o_2}^{C_1^{(2)}} u_{o_2}(7) - d_1 \geq \hat{w}_{i_1}^{C_1^{(2)}} u_{i_1}(9) + \hat{w}_{o_1}^{C_1^{(2)}} u_{o_1}(5) + \hat{w}_{i_2}^{C_1^{(2)}} u_{i_2}(8) + \hat{w}_{o_2}^{C_1^{(2)}} u_{o_2}(7) - C(1 - b_1) \\
 & \hat{w}_{i_1}^{C_1^{(2)}} u_{i_1}(2) + \hat{w}_{o_1}^{C_1^{(2)}} u_{o_1}(8) + \hat{w}_{i_2}^{C_1^{(2)}} u_{i_2}(5) + \hat{w}_{o_2}^{C_1^{(2)}} u_{o_2}(5) - d_1 \geq \hat{w}_{i_1}^{C_1^{(2)}} u_{i_1}(9) + \hat{w}_{o_1}^{C_1^{(2)}} u_{o_1}(5) + \hat{w}_{i_2}^{C_1^{(2)}} u_{i_2}(8) + \hat{w}_{o_2}^{C_1^{(2)}} u_{o_2}(7) - C(1 - b_2) \\
 & \hat{w}_{i_1}^{C_1^{(2)}} u_{i_1}(5) + \hat{w}_{o_1}^{C_1^{(2)}} u_{o_1}(6) + \hat{w}_{i_2}^{C_1^{(2)}} u_{i_2}(7) + \hat{w}_{o_2}^{C_1^{(2)}} u_{o_2}(6) - d_1 \geq \hat{w}_{i_1}^{C_1^{(2)}} u_{i_1}(9) + \hat{w}_{o_1}^{C_1^{(2)}} u_{o_1}(5) + \hat{w}_{i_2}^{C_1^{(2)}} u_{i_2}(8) + \hat{w}_{o_2}^{C_1^{(2)}} u_{o_2}(7) - C(1 - b_3) \\
 & \hat{w}_{i_1}^{C_1^{(2)}} u_{i_1}(1) + \hat{w}_{o_1}^{C_1^{(2)}} u_{o_1}(2) + \hat{w}_{i_2}^{C_1^{(2)}} u_{i_2}(9) + \hat{w}_{o_2}^{C_1^{(2)}} u_{o_2}(8) - d_1 \geq \hat{w}_{i_1}^{C_1^{(2)}} u_{i_1}(9) + \hat{w}_{o_1}^{C_1^{(2)}} u_{o_1}(5) + \hat{w}_{i_2}^{C_1^{(2)}} u_{i_2}(8) + \hat{w}_{o_2}^{C_1^{(2)}} u_{o_2}(7) - C(1 - b_4) \\
 & b_1 + b_2 + b_3 + b_4 = 1 \\
 & b_1, b_2, b_3, b_4 \in \{0, 1\} \\
 & d_1 \geq 0, \\
 & \mathcal{W}^{C_1^{(2)}}
 \end{aligned} \right\} \tag{A2}$$

The following model allows us to find the minimal (best) efficiency rank for DMU_1 :

$$\text{Minimize } 1 + b_2 + b_3 + b_4$$

s.t.

$$\left. \begin{aligned}
 & \hat{w}_{i_1}^{C_1^{(2)}} u_{i_1}(2) + \hat{w}_{o_1}^{C_1^{(2)}} u_{o_1}(8) + \hat{w}_{i_2}^{C_1^{(2)}} u_{i_2}(5) + \hat{w}_{o_2}^{C_1^{(2)}} u_{o_2}(5) - (\hat{w}_{i_1}^{C_1^{(2)}} u_{i_1}(9) + \hat{w}_{o_1}^{C_1^{(2)}} u_{o_1}(5) + \hat{w}_{i_2}^{C_1^{(2)}} u_{i_2}(8) + \hat{w}_{o_2}^{C_1^{(2)}} u_{o_2}(7)) \leq Cb_2 \\
 & \hat{w}_{i_1}^{C_1^{(2)}} u_{i_1}(5) + \hat{w}_{o_1}^{C_1^{(2)}} u_{o_1}(6) + \hat{w}_{i_2}^{C_1^{(2)}} u_{i_2}(7) + \hat{w}_{o_2}^{C_1^{(2)}} u_{o_2}(6) - (\hat{w}_{i_1}^{C_1^{(2)}} u_{i_1}(9) + \hat{w}_{o_1}^{C_1^{(2)}} u_{o_1}(5) + \hat{w}_{i_2}^{C_1^{(2)}} u_{i_2}(8) + \hat{w}_{o_2}^{C_1^{(2)}} u_{o_2}(7)) \leq Cb_3 \\
 & \hat{w}_{i_1}^{C_1^{(2)}} u_{i_1}(1) + \hat{w}_{o_1}^{C_1^{(2)}} u_{o_1}(2) + \hat{w}_{i_2}^{C_1^{(2)}} u_{i_2}(9) + \hat{w}_{o_2}^{C_1^{(2)}} u_{o_2}(8) - (\hat{w}_{i_1}^{C_1^{(2)}} u_{i_1}(9) + \hat{w}_{o_1}^{C_1^{(2)}} u_{o_1}(5) + \hat{w}_{i_2}^{C_1^{(2)}} u_{i_2}(8) + \hat{w}_{o_2}^{C_1^{(2)}} u_{o_2}(7)) \leq Cb_4 \\
 & b_2, b_3, b_4 \in \{0, 1\} \\
 & \mathcal{W}^{C_1^{(2)}}
 \end{aligned} \right\} \tag{A3}$$

The maximal (worst) efficiency rank for DMU_1 can be determined by solving the following model:

$$\text{Maximize } 1 + b_2 + b_3 + b_4$$

s.t.

$$\left. \begin{aligned}
 & \hat{w}_{i_1}^{C_1^{(2)}} u_{i_1}(9) + \hat{w}_{o_1}^{C_1^{(2)}} u_{o_1}(5) + \hat{w}_{i_2}^{C_1^{(2)}} u_{i_2}(8) + \hat{w}_{o_2}^{C_1^{(2)}} u_{o_2}(7) - (\hat{w}_{i_1}^{C_1^{(2)}} u_{i_1}(2) + \hat{w}_{o_1}^{C_1^{(2)}} u_{o_1}(8) + \hat{w}_{i_2}^{C_1^{(2)}} u_{i_2}(5) + \hat{w}_{o_2}^{C_1^{(2)}} u_{o_2}(5)) \leq C(1 - b_2) \\
 & \hat{w}_{i_1}^{C_1^{(2)}} u_{i_1}(9) + \hat{w}_{o_1}^{C_1^{(2)}} u_{o_1}(5) + \hat{w}_{i_2}^{C_1^{(2)}} u_{i_2}(8) + \hat{w}_{o_2}^{C_1^{(2)}} u_{o_2}(7) - (\hat{w}_{i_1}^{C_1^{(2)}} u_{i_1}(5) + \hat{w}_{o_1}^{C_1^{(2)}} u_{o_1}(6) + \hat{w}_{i_2}^{C_1^{(2)}} u_{i_2}(7) + \hat{w}_{o_2}^{C_1^{(2)}} u_{o_2}(6)) \leq C(1 - b_3) \\
 & \hat{w}_{i_1}^{C_1^{(2)}} u_{i_1}(9) + \hat{w}_{o_1}^{C_1^{(2)}} u_{o_1}(5) + \hat{w}_{i_2}^{C_1^{(2)}} u_{i_2}(8) + \hat{w}_{o_2}^{C_1^{(2)}} u_{o_2}(7) - (\hat{w}_{i_1}^{C_1^{(2)}} u_{i_1}(1) + \hat{w}_{o_1}^{C_1^{(2)}} u_{o_1}(2) + \hat{w}_{i_2}^{C_1^{(2)}} u_{i_2}(9) + \hat{w}_{o_2}^{C_1^{(2)}} u_{o_2}(8)) \leq C(1 - b_4) \\
 & b_2, b_3, b_4 \in \{0, 1\} \\
 & \mathcal{W}^{C_1^{(2)}}
 \end{aligned} \right\} \tag{A4}$$

When referring to pairwise preference relations, we consider an ordered pair (DMU_1, DMU_2) . The truth of the necessary preference relation can be verified by solving the following model:

$$\text{Minimize } d_{1,2}$$

s.t.

$$\left. \begin{aligned}
 & \hat{w}_{i_1}^{C_1^{(2)}} u_{i_1}(9) + \hat{w}_{o_1}^{C_1^{(2)}} u_{o_1}(5) + \hat{w}_{i_2}^{C_1^{(2)}} u_{i_2}(8) + \hat{w}_{o_2}^{C_1^{(2)}} u_{o_2}(7) - (\hat{w}_{i_1}^{C_1^{(2)}} u_{i_1}(2) + \hat{w}_{o_1}^{C_1^{(2)}} u_{o_1}(8) + \hat{w}_{i_2}^{C_1^{(2)}} u_{i_2}(5) + \hat{w}_{o_2}^{C_1^{(2)}} u_{o_2}(5)) \leq d_{1,2} \\
 & \mathcal{W}^{C_1^{(2)}}
 \end{aligned} \right\} \tag{A5}$$

If the minimal distance $d_{1,2}$ is not lesser than zero, then $DMU_1 \succ_E^{N, C_1^{(2)}} DMU_2$. The following model allows us to verify the truth of the possible preference relation:

$$\text{Maximize } d_{1,2}$$

s.t.

$$\left. \begin{aligned} & \hat{w}_{i_1}^{C_1^{(2)}} u_{i_1}(9) + \hat{w}_{o_1}^{C_1^{(2)}} u_{o_1}(5) + \hat{w}_{i_2}^{C_1^{(2)}} u_{i_2}(8) + \hat{w}_{o_2}^{C_1^{(2)}} u_{o_2}(7) - (\hat{w}_{i_1}^{C_1^{(2)}} u_{i_1}(2) + \hat{w}_{o_1}^{C_1^{(2)}} u_{o_1}(8) + \hat{w}_{i_2}^{C_1^{(2)}} u_{i_2}(5) + \hat{w}_{o_2}^{C_1^{(2)}} u_{o_2}(5)) \geq d_{1,2} \\ & \mathcal{W}^{C_1^{(2)}}. \end{aligned} \right\} \quad (A6)$$

If the maximal value of $d_{1,2}$ is greater or equal to zero, then $DMU_1 \underset{E}{\succ}^{P, C_1^{(2)}} DMU_2$.

Appendix C. Computation of Stochastic Acceptability Indices

For illustrative purposes, we present the steps of computing the stochastic acceptability indices for the exemplary DMUs from Appendix B considering the root category ($C_1^{(2)}$). Firstly, we use the hit-and-run algorithm to generate the weight samples for all children of $C_1^{(2)}$. Table A2 presents the five example samples. Let us emphasize that to obtain reliable estimates of stochastic acceptabilities, in practice, one uses a few thousand such samples. Secondly, we compute a value-based efficiency score for each considered DMU (see Table A3). Based on the efficiency scores, we determine the distance of each DMU_o to the best one as the difference between the maximal efficiency score obtained by any DMU_k for a given sample and the efficiency score of DMU_o . For example, for sample 1 and DMU_3 , such distance equals $0.773 - 0.542 = 0.231$. The efficiency rank of DMU_o is equal to the number of other DMUs, for which the efficiency score is better than that of DMU_o (for a given sample), increased by one. For example, for sample 1, DMU_2 and DMU_3 attained an efficiency score greater than DMU_4 , ranking it 3rd. The distances to the best DMU and efficiency ranks for all exemplary DMUs and samples are presented in Table A3.

Table A2. Five example weight vectors obtained with Monte Carlo simulation.

Sample	1	2	3	4	5
w_{i_1}	0.33	0.52	0.56	0.81	0.7
w_{o_1}	0.67	0.48	0.44	0.19	0.3
w_{i_2}	0.71	0.53	0.46	0.68	0.46
w_{o_2}	0.29	0.47	0.54	0.32	0.54
$w_{C_1^{(1)}}$	0.7	0.68	0.75	0.53	0.69
$w_{C_2^{(1)}}$	0.3	0.32	0.25	0.47	0.31

Having determined the distances to the best DMU and efficiency ranks, we compute the stochastic indices. $DAII$ is the share of the weight vectors for which the distance to the best DMU is within a given bucket. For illustrative purposes, we use only four buckets. For example, there are four out of five samples for which the distance of DMU_4 to the best unit is in the interval $b_1 = [0, 0.25]$, so $DAII(DMU_4, [0, 0.25]) = 0.8$. Similarly, we calculate $ERAI$ s as the share of samples for which a given DMU is ranked r -th. For example, DMU_3 is ranked 2nd only for sample 1, so $ERAI(DMU_3, 2) = 0.2$. Finally, $PEOI$ for an ordered pair of units is computed as the share of samples for which the first DMU is at least as efficient as the other. For example, DMU_4 is not worse than DMU_3 for all samples except the first one, so $PEOI(DMU_4, DMU_3) = 0.8$. The results obtained by individual samples are averaged, providing the estimates of the expected distances Ed to the best DMU and expected efficiency ranks ER (see Table A3).

Table A3. Efficiency scores E , distances d , and ranks R for the considered DMUs obtained with five sampled weight vectors.

Sample	Result	DMU ₁	DMU ₂	DMU ₃	DMU ₄
1	E	0.378	0.773	0.542	0.397
	d	0.395	0	0.231	0.377
	R	4	1	2	3
2	E	0.327	0.714	0.504	0.508
	d	0.387	0	0.209	0.206
	R	4	1	3	2
3	E	0.307	0.715	0.504	0.548
	d	0.408	0	0.211	0.167
	R	4	1	3	2
4	E	0.281	0.744	0.497	0.561
	d	0.462	0	0.247	0.183
	R	4	1	3	2
5	E	0.276	0.695	0.485	0.605
	d	0.419	0	0.210	0.090
	R	4	1	3	2
	Ed	0.251	0.202	0.304	0.377
	ER	4.0	1.0	2.8	2.2

References

- Charnes, A.; Cooper, W.W.; Rhodes, E. Measuring the efficiency of decision making units. *Eur. J. Oper. Res.* **1978**, *2*, 429–444. [\[CrossRef\]](#)
- Liu, J.S.; Lu, L.Y.; Lu, W.M.; Lin, B.J. A survey of DEA applications. *Omega* **2013**, *41*, 893–902. [\[CrossRef\]](#)
- Krejnus, M.; Stofkova, J.; Stofkova, K.R.; Binasova, V. The Use of the DEA Method for Measuring the Efficiency of Electronic Public Administration as Part of the Digitization of the Economy and Society. *Appl. Sci.* **2023**, *13*, 3672. [\[CrossRef\]](#)
- Wang, L.; Lu, F.; Han, B.; Zhang, Q.; Zhang, C. Simulation-Based Optimization of Transport Efficiency of an Urban Rail Transit Network. *Appl. Sci.* **2023**, *13*, 1471. [\[CrossRef\]](#)
- Emrouznejad, A.; Parker, B.R.; Tavares, G. Evaluation of research in efficiency and productivity: A survey and analysis of the first 30 years of scholarly literature in DEA. *Socio-Econ. Plan. Sci.* **2008**, *42*, 151–157. [\[CrossRef\]](#)
- Cooper, W.W.; Seiford, L.M.; Zhu, J. Data envelopment analysis: History, models, and interpretations. In *Handbook on Data Envelopment Analysis*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 1–39.
- Charnes, A.; Cooper, W.W.; Golany, B.; Seiford, L.; Stutz, J. Foundations of data envelopment analysis for Pareto-Koopmans efficient empirical production functions. *J. Econom.* **1985**, *30*, 91–107. [\[CrossRef\]](#)
- Gouveia, M.C.; Dias, L.C.; Antunes, C.H. Additive DEA based on MCDA with imprecise information. *J. Oper. Res. Soc.* **2008**, *59*, 54–63. [\[CrossRef\]](#)
- de Almeida, P.; Dias, L. Value-based DEA models: Application-driven developments. *J. Oper. Res. Soc.* **2012**, *63*, 16–27. [\[CrossRef\]](#)
- Keeney, R.L.; Raiffa, H.; Meyer, R.F. *Decisions with Multiple Objectives: Preferences and Value Trade-Offs*; Cambridge University Press: Cambridge, UK, 1993.
- Alidrisi, H. DEA-Based PROMETHEE II Distribution-Center Productivity Model: Evaluation and Location Strategies Formulation. *Appl. Sci.* **2021**, *11*, 9567. [\[CrossRef\]](#)
- Bagherikahvarin, M.; De Smet, Y. A ranking method based on DEA and PROMETHEE II (a rank based on DEA & PR.II). *Measurement* **2016**, *89*, 333–342.
- Saaty, T.L. What is the Analytic Hierarchy Process? In *Mathematical Models for Decision Support*; Springer: Berlin/Heidelberg, Germany, 1988; pp. 109–121.
- Corrente, S.; Greco, S.; Słowiński, R. Multiple Criteria Hierarchy Process in Robust Ordinal Regression. *Decis. Support Syst.* **2012**, *53*, 660–674. [\[CrossRef\]](#)
- Corrente, S.; Greco, S.; Słowiński, R. Multiple Criteria Hierarchy Process with ELECTRE and PROMETHEE. *Omega* **2013**, *41*, 820–846. [\[CrossRef\]](#)
- Del Vasto-Terrientes, L.; Valls, A.; Slowinski, R.; Zielniewicz, P. ELECTRE-III-H: An outranking-based decision aiding method for hierarchically structured criteria. *Expert Syst. Appl.* **2015**, *42*, 4910–4926. [\[CrossRef\]](#)
- Meng, W.; Zhang, D.; Qi, L.; Liu, W. Two-level DEA approaches in research evaluation. *Omega* **2008**, *36*, 950–957. [\[CrossRef\]](#)
- Kao, C. A linear formulation of the two-level DEA model. *Omega* **2008**, *36*, 958–962. [\[CrossRef\]](#)
- Shen, Y.; Hermans, E.; Ruan, D.; Wets, G.; Brijs, T.; Vanhoof, K. A generalized multiple layer data envelopment analysis model for hierarchical structure assessment: A case study in road safety performance evaluation. *Expert Syst. Appl.* **2011**, *38*, 15262–15272. [\[CrossRef\]](#)
- Pakkar, M.S. An integrated approach based on DEA and AHP. *Comput. Manag. Sci.* **2015**, *12*, 153–169. [\[CrossRef\]](#)

21. Pakkar, M.S. A hierarchical aggregation approach for indicators based on data envelopment analysis and analytic hierarchy process. *Systems* **2016**, *4*, 6. [[CrossRef](#)]
22. Amini, M.R.; Azar, A.; Eskandari, H.; Wanke, P.F. A generalized fuzzy Multiple-Layer NDEA: An application to performance-based budgeting. *Appl. Soft Comput.* **2021**, *100*, 106984. [[CrossRef](#)]
23. Kadziński, M.; Labijak, A.; Napieraj, M. Integrated framework for robustness analysis using ratio-based efficiency model with application to evaluation of Polish airports. *Omega* **2017**, *67*, 1–18. [[CrossRef](#)]
24. Labijak-Kowalska, A.; Kadziński, M.; Spychała, I.; Dias, L.C.; Fiallos, J.; Patrick, J.; Michalowski, W.; Farion, K. Performance evaluation of emergency department physicians using robust value-based additive efficiency model. *Int. Trans. Oper. Res.* **2023**, *30*, 503–544. [[CrossRef](#)]
25. Miszczyńska, K.; Miszczyński, P.M. Measuring the efficiency of the healthcare sector in Poland—A window-DEA evaluation. *Int. J. Product. Perform. Manag.* **2022**, *71*, 2743–2770. [[CrossRef](#)]
26. Zakowska, I.; Godycki-Cwirko, M. Data envelopment analysis applications in primary health care: A systematic review. *Fam. Pract.* **2019**, *37*, 147–153. [[CrossRef](#)] [[PubMed](#)]
27. Ciomek, K.; Kadziński, M. Polyrun: A Java library for sampling from the bounded convex polytopes. *SoftwareX* **2021**, *13*, 100659. [[CrossRef](#)]
28. Labijak-Kowalska, A.; Kadziński, M. Experimental comparison of results provided by ranking methods in Data Envelopment Analysis. *Expert Syst. Appl.* **2021**, *173*, 114739. [[CrossRef](#)]
29. Kozierkiewicz, A.; Natkaniec, M.; Megas, B.; Gilewski, D.; Ignatowicz, M.; Waško, B. *Indeks Sprawności Ochrony Zdrowia 2018*; Technical Report; PricewaterhouseCoopers: Warsaw, Poland, 2019. (In Polish)
30. Sexton, T.R.; Silkman, R.H.; Hogan, A.J. Data envelopment analysis: Critique and extensions. *New Dir. Program Eval.* **1986**, *1986*, 73–105. [[CrossRef](#)]
31. Andersen, P.; Petersen, N.C. A procedure for ranking efficient units in data envelopment analysis. *Manag. Sci.* **1993**, *39*, 1261–1264. [[CrossRef](#)]
32. Kendall, M.G. A new measure of rank correlation. *Biometrika* **1938**, *30*, 81–93. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Extended abstract in Polish

Nowe Kierunki w Odpornej Granicznej Analizie Danych

Wprowadzenie

Graniczna Analiza Danych (ang. Data Envelopment Analysis (DEA)) jest metodą pozwalającą na ocenę względnej efektywności jednostek decyzyjnych, które pobierają wiele nakładów (wejść) i produkują wiele efektów (wyjść). W oryginalnym sformułowaniu metody, efektywność jednostki określana jest jako stosunek pomiędzy wirtualnym efektem a wirtualnym nakładem, będącymi sumami ważonymi istniejących nakładów i efektów. Celem metody DEA jest identyfikacja zbioru jednostek, które działają efektywnie. Jest on uzyskiwany poprzez znalezienie wektora wag skojarzonych z nakładami i efektami, dla którego badana jednostka osiąga najwyższą wartość miary efektywności.

W Granicznej Analizie Danych wyróżnia się trzy podstawowe modele efektywności: CCR, BCC oraz model addytywny. Pierwsze dwa są modelami ilorazowymi, w których rozróżnia się orientację na nakłady i na efekty. Ostatni opiera się na wyznaczeniu odległości L_1 badanej jednostki od granicy efektywności. W tym przypadku orientacje na nakłady i efekty zostały połączone w jeden wspólny model. Choć takie powiązanie obu orientacji jest często pożądanym efektem, to oryginalne sformułowanie modelu addytywnego posiada kilka wad. Po pierwsze, w modelu addytywnym istnieje problem skali, tj. projekcje jednostek nieefektywnych na granicę efektywności zależą, w znacznym stopniu, od zakresów wartości poszczególnych czynników (wejść i wyjść). Po drugie, miara efektywności w tym modelu nie ma intuicyjnej interpretacji. Rozwiązaniem powyższych problemów było wprowadzenie nowego modelu efektywności, tj. modelu addytywnego opartego na funkcjach wartości, (ang. Value-based additive DEA (VDEA)), który czerpie inspirację z wieloatrybutowej teorii użyteczności. W tym modelu, wejściom i wyjściom przypisuje się cząstkowe funkcje wartości, które następnie agregowane są z użyciem funkcji addytywnej (sumy ważonej).

Niezależnie od wybranego modelu efektywności, ocena jednostek decyzyjnych oparta jest zawsze tylko na jednym, najbardziej korzystnym, wektorze wag dla badanej jednostki. Ponadto, wektory wag uzyskane dla różnych jednostek są różne, co budzi wątpliwości co do merytorycznej poprawności porównań między nimi. Dodatkowo, klasyczne podejście

w DEA pozwala tylko na wskazanie jednostek efektywnych, nie dostarczając przy tym narzędzi pozwalających na ich rozróżnienie.

Opisane powyżej problemy były punktem wyjścia do dalszego rozwoju metodologii DEA. Po pierwsze, w literaturze zaproponowano wiele metod, które pozwalają na porównanie jednostek efektywnych między sobą oraz stworzenie pełnego rankingu jednostek decyzyjnych, spośród których najbardziej popularne są super-efektywność oraz efektywność krzyżowa. Ponadto, wprowadzono możliwość uwzględnienia informacji preferencyjnej w modelu, co pozwala na ograniczenie przestrzeni dopuszczalnych wektorów wag. Żadna z powyższych metod nie pozwala jednak na ocenę jednostki z uwzględnieniem całego dopuszczalnego spektrum wektorów wag. Badania przeprowadzone w ramach niniejszej rozprawy skupiają się na opracowaniu spójnego zestawu metod, pozwalających na ocenę odporności jednostek decyzyjnych, tj. ich efektywności w oparciu o wszystkie dopuszczalne wektory wag. W kolejnych rozdziałach zaproponowane rozwiązania zostały opisane w zwięzłej formie.

1 Metody badania odporności dla Granicznej Analizy Danych

W niniejszej rozprawie zaproponowany został spójny zestaw metod, które pozwalają na badanie odporności jednostek decyzyjnych w oparciu o pełen zakres wektorów wag dla dwóch różnych modeli efektywności: ilorazowego oraz opartego na funkcjach wartości. Zaproponowane metody można podzielić na dwie grupy, które wzajemnie się uzupełniają. Pierwsza z nich oparta jest na programowaniu matematycznym i pozwala na wyznaczenie skrajnych wartości efektywności, odległości od najlepszej jednostki oraz występowanie koniecznych i możliwych relacji preferencji dla par jednostek. Druga wykorzystuje symulację Monte Carlo do wyznaczenia indeksów akceptowalności reprezentujących rozkłady poszczególnych miar lub wyników.

Równoczesna analiza wyników uzyskanych w obu podejściach jest korzystna z kilku względów. Z jednej strony, metody dokładne pozwalają na precyzyjną ocenę tego, w jakim zakresie znajdują się wartości efektywności dla poszczególnych jednostek z uwzględnieniem wszystkich wektorów wag. Uzyskane zakresy są jednak w wielu przypadkach bardzo szerokie. Dodatkowo, uzyskane skrajne wartości występują bardzo rzadko, tylko dla pojedynczych, szczególnych wektorów wag. Podobnie, w wielu problemach relacja możliwej preferencji występuje dla większości par jednostek decyzyjnych, podczas gdy konieczna preferencja jest relacją rzadką. Powyższa własność czyni większość par jednostek nieporównywalną. W takich sytuacjach analiza stochastyczna pozwala na uzyskanie dodatkowej informacji na temat rozkładów badanych miar, tj. wartości miary efektywności, odległości do najlepszej jednostki, pozycji w rankingu oraz prawdopodobieństwa preferencji jednej jednostki nad inną. Z drugiej strony, same indeksy akceptowalności również nie dostarczają wszystkich potrzebnych informacji. Mogą one być wyznaczone jedynie w sposób przybliżony. W szczególności, szansa, że wylosowany zostanie wektor wag odpowiadający najwyższej lub najniższej możliwej wartości efektywności dla badanej jednostki jest niewielka. Analogicznie, szacowana wartość indeksu akceptowalności relacji przewyższania dla pewnej pary jednostek może być równa 1, jednak nie świadczy to o koniecznej preferencji w obrębie tej pary. Powyższe cechy świadczą o zasadności zestawienia ze sobą obu typów wyników.

Aby uzyskać stochastyczne indeksy akceptowalności dla problemów z ilorazowym modelem efektywności zastosowano poniższy algorytm. Jako, że przestrzeń wektorów wag jest nieograniczona, wprowadzono dodatkowe ograniczenia, które normalizują otrzymane wektory wag w taki sposób, że suma wag, zarówno dla nakładów jak i dla efektów musi być równa 1. Następnie, zastosowano algorytm Hit-And-Run, aby uzyskać zbiór próbek wektorów wag. W niniejszej pracy do próbkowania użyto rozkładu jednorodnego. Bazując na wygenerowanym zestawie wag obliczone zostają wartości efektywności dla wszystkich jednostek. W ostatnim etapie otrzymane wartości efektywności zostają podzielone przez wartość uzyskaną przez najlepszą jednostkę. Taki zabieg zapewnia, że najlepsza jednostka uzyskuje efektywność równą 1. Uzyskane w taki sposób wartości efektywności pozwalają na wyznaczenie rozkładów miary efektywności oraz pozycji w rankingu efektywności. Dodatkowo, estymowane są również prawdopodobieństwa występowania preferencji dla par jednostek.

Rozważone zostały trzy różne perspektywy oceny efektywności jednostek decyzyjnych: wartości miary efektywności (lub ich odległość do najlepszej jednostki), rankingi jednostek pod względem efektywności oraz ich porównania parami. Poniżej, poszczególne perspektywy zostały krótko opisane w odniesieniu do modelu ilorazowego.

Wartości miary efektywności

Z punktu widzenia miary efektywności analiza odporności polega na wyznaczeniu skrajnych (maksymalnych i minimalnych) wartości miary efektywności dla każdej jednostki. Maksymalną miarę efektywności dla badanej jednostki można uzyskać rozwiązując oryginalnie zaproponowany w metodzie DEA model programowania liniowego (model CCR). W niniejszej rozprawie zaproponowany został model mieszanego programowania całkowitoliczbowego, który pozwala na znalezienie najmniej korzystnego scenariusza dla badanej jednostki. W zaproponowanym modelu minimalizowana jest miara efektywności badanej jednostki z uwzględnieniem ograniczenia, że przynajmniej jedna z jednostek w zbiorze pozostaje efektywna, tj. osiąga efektywną równą jeden.

Rozkład miary efektywności w dopuszczalnej przestrzeni wektorów wag wyznaczany jest poprzez oszacowanie indeksów akceptowalności dla przedziałów efektywności (ang. Efficiency Acceptability Interval Indices (EAIIs)). Indeks ten dla pewnej jednostki DMU_o oraz przedziału b_i definiowany jest jako stosunek liczby wektorów wag, dla których jednostka DMU_o osiąga efektywność w przedziale b_i do liczby wszystkich dopuszczalnych wektorów wag. Przedziały $b_i, i = 1, 2, \dots, K$ muszą być rozłączne i pokrywać całą możliwą przestrzeń miary efektywności, tj. ich suma musi być równa przedziałowi $[0, 1]$. Suma wartości EAII dla danej jednostki jest zawsze równa jeden.

Dodatkowo, analiza z użyciem symulacji Monte Carlo pozwala na wyznaczenie dodatkowych miar opisujących efektywność jednostek, takich jak skrajne wartości efektywności uzyskane przy pomocy próbkowania oraz oczekiwana wartość efektywności.

Rankingi jednostek pod względem efektywności.

Drugą perspektywą analizy jednostek decyzyjnych jest ocena danej jednostki pod względem pozycji, którą zajmuje w rankingu efektywności. W tej perspektywie wyznaczane są skrajne pozycje, jakie może osiągać dana jednostka dla jakiegokolwiek dopuszczalnego wektora wag. W ramach niniejszej pracy zaproponowano modele programowania linio-

wego pozwalające na wyznaczenie tych skrajnych pozycji. Oba zaproponowane modele wykorzystują mieszane programowanie całkowitoliczbowe. W obu przypadkach model opiera się na ustaleniu wartości efektywności badanej jednostki na 1. W przypadku modelu znajdującego najlepszą (minimalną) pozycję danej jednostki minimalizuje się sumę zmiennych binarnych odpowiadających jednostkom, które równocześnie osiągają efektywność większą od badanej jednostki. Analogicznie, w modelu znajdującym najgorszą pozycję w rankingu dla danej jednostki, suma zmiennych binarnych odpowiadająca jednostkom nie gorszym niż badana jednostka jest maksymalizowana.

Aby uzyskać rozkład pozycji w rankingu dla badanej jednostki wyznaczone zostały wartości indeksy akceptowalności dla rankingów efektywności (ang. Efficiency Rank Acceptability Indices (ERAIs)) zdefiniowane jako udział wektorów wag, dla których badana jednostka osiąga daną pozycję w rankingu efektywności. Dla każdej jednostki suma wartości ERAI dla wszystkich pozycji jest równa 1. Oprócz rozkładu pozycji w rankingach, dla każdej jednostki wyznaczana jest również wartość oczekiwana pozycji w rankingu efektywności, jako średnia arytmetyczna pozycji uzyskanych we wszystkich próbkach.

Porównania parami jednostek decyzyjnych.

Ostatnia perspektywa analizy rozważana w ramach niniejszej rozprawy skupia się na porównaniach parami jednostek decyzyjnych. W tym przypadku zdefiniowano dwie relacje preferencji pomiędzy jednostkami decyzyjnymi. Relacja możliwej preferencji (\succsim_E^P) dla pary jednostek zachodzi wtedy, gdy istnieje choć jeden dopuszczalny wektor wag, dla którego pierwsza jednostka osiąga efektywność nie gorszą niż druga. Relacja koniecznej preferencji (\succsim_E^N) zachodzi dla pary jednostek, gdy pierwsza z jednostek osiąga efektywność większą lub równą efektywności drugiej jednostki dla wszystkich dopuszczalnych wektorów wag.

Aby zweryfikować występowanie relacji możliwej preferencji dla danej pary jednostek decyzyjnych, zaproponowano model programowania liniowego, w którym maksymalizowana jest wartość efektywności pierwszej jednostki przy założeniu, że efektywność drugiej z nich jest równa 1. Jeśli otrzymana optymalna efektywność jest większa lub równa 1, wtedy pierwsza jednostka jest możliwie preferowana nad drugą. Taki sam model, po zmianie kierunku optymalizacji, pozwala na sprawdzenie czy dana jednostka jest koniecznie preferowana nad inną. Jeśli minimalna efektywność uzyskana w opisywanym modelu jest nie mniejsza niż 1, wtedy pierwsza jednostka z badanej pary jest koniecznie preferowana nad drugą.

Ponadto, dla pary jednostek decyzyjnych zdefiniowano indeks akceptowalności dla relacji przewyższania pod względem efektywności (ang. Pairwise Efficiency Outranking Index (PEOI)) jako liczbę dopuszczalnych wektorów wag, dla których pierwsza z badanych jednostek osiąga efektywność nie gorszą niż druga w stosunku do wszystkich wektorów wag.

Metody badania odporności dla modelu addytywnego opartego na funkcjach wartości.

Opisane powyżej miary mogą zostać zastosowane, w analogiczny sposób, dla problemów, w których zastosowano model efektywności oparty na funkcjach wartości. Podobnie jak dla modelu ilorazowego, rozważone zostały trzy perspektywy analizy efektywności jed-

nostek: wartości miary efektywności, pozycje w rankingach oraz porównania parami jednostek decyzyjnych.

W kontekście modelu opartego o funkcje wartości perspektywa wartości miary efektywności może uwzględniać zarówno bezwzględne wartości efektywności jak i odległość badanej jednostki od najlepszej. W niniejszej rozprawie rozważono obie te miary. Po pierwsze, wykorzystano oryginalne sformułowanie modelu VDEA do wyznaczenia najlepszej (minimalnej) odległości badanej jednostki od najlepszej. Następnie, zaproponowano model matematyczny pozwalający na wyznaczenie najgorszej możliwej odległości badanej jednostki od jednostki najlepszej.

Rozważając bezwzględne wartości efektywności w modelu opartym o funkcje wartości, zaproponowano metodę, która pozwala na określenie skrajnych wartości efektywności dla badanej jednostki. W tym przypadku optymalizowana jest miara efektywności z zapewnieniem zachowania zdefiniowanych ograniczeń na wagi nakładów i efektów.

Analiza stochastyczna dla modelu VDEA prowadzona jest w sposób analogiczny do tej prowadzonej dla modelu ilorazowego, z tą różnicą, że w modelu opartym na funkcjach wartości suma wszystkich wag, przypisanych do wejść i wyjść musi być równa 1. Po wylosowaniu próbek wag wyznaczana jest wartość efektywności dla wszystkich jednostek. Z uwagi na zastosowanie w rozważanym modelu cząstkowych funkcji wartości oraz normalizacji wag, osiągane wartości efektywności zawierają się zawsze w przedziale $[0 - 1]$. W opozycji do modelu ilorazowego, dodatkowa normalizacja nie jest potrzebna.

Bazując na otrzymanych wartościach efektywności, estymowane są indeksy akceptowalności (EAII, PEOI, ERAI) w sposób analogiczny jak dla modelu ilorazowego. Dodatkowo, w przypadku modelu VDEA, wyznaczany jest indeks akceptowalności dla przedziałów odległości do najlepszej jednostki w sposób analogiczny do wyznaczania indeksów EAII. Ponadto, możliwe jest również oszacowanie wartości oczekiwanej tej odległości dla każdej jednostki decyzyjnej.

Zależności między wynikami dokładnymi a indeksami akceptowalności.

W rozprawie przedstawione zostały zależności pomiędzy wynikami otrzymanymi przy użyciu programowania liniowego oraz szacowanymi indeksami akceptowalności. Poniżej przedstawiono przykładowe z nich. Dla każdej jednostki DMU_o :

- jeżeli cały przedział b_i leży poza zakresem wyznaczonym przez skrajne wartości efektywności dla DMU_o , to $EAII(DMU_o, b_i) = 0$;
- przedział wyznaczony przez skrajne wartości efektywności otrzymane przy pomocy próbkowania zawsze zawiera się w przedziale wyznaczonym przez rzeczywiste skrajne wartości efektywności;
- jeżeli dana jednostka jest koniecznie preferowana nad inną jednostkę to PEOI dla tej pary jednostek jest zawsze równe 1;
- suma wartości ERAI dla wszystkich pozycji z przedziału wyznaczonego przez skrajne pozycje musi być równa 1.

Relacje przeciwnie nie są prawdziwe ze względu na losowy charakter wyznaczanych indeksów akceptowalności. Na przykład, jeśli PEOI uzyskane dla pewnej pary jednostek

decyzyjnych jest równe 1, nie oznacza, że preferencja dla tej pary występuje dla wszystkich wektorów wag. Oznacza to tylko tyle, że dla każdej wylosowanej próbki jedna z jednostek osiągała efektywność nie gorszą niż druga. Jednak istnieje możliwość, że opisana preferencja nie zachodzi dla pewnego wektora wag, który nie został wylosowany.

Wpływ przyrostowego definiowania ograniczeń wag na wyniki badania odporności.

W rozprawie omówiony został także wpływ przyrostowej definicji ograniczeń wag na wyniki badania odporności. Rozważono sytuację, w której analiza odporności tego samego zbioru jednostek została przeprowadzona kilkakrotnie, za każdym razem rozszerzając ograniczenia na wagi. W rezultacie przestrzeń możliwych wektorów wag, w kolejnych iteracjach ulegała zawężeniu. W takim przypadku przedziały wyznaczone przez skrajne wartości efektywności, odległości do najlepszej jednostki oraz pozycje w rankingu efektywności zawężają się w kolejnych iteracjach. Wartości minimalne są, w kolejnych iteracjach, nierosnące, podczas gdy wartości maksymalne mogą tylko maleć (lub pozostać niezmienione). Ponadto, relacja koniecznej preferencji, w kolejnych iteracjach, ulega wzbogaceniu, tj. zbiór par jednostek, dla których ta relacja zachodzi w iteracji t jest nadzbiorem takiego zbioru dla iteracji $t - 1$. Odwrotną zależność można zauważyć dla relacji możliwych preferencji. W kolejnych iteracjach zbiór par jednostek, dla których zachodzi relacja możliwej preferencji jest coraz mniejszy.

2 Badanie odporności jednostek decyzyjnych dla problemów z nieprecyzyjnymi danymi

W rzeczywistych problemach zebranie precyzyjnych danych jest często niemożliwe lub bardzo kosztowne. Z tego powodu, w ramach niniejszej rozprawy, metody badania odporności zostały rozszerzone i zaadaptowane do problemów z nieprecyzyjnymi danymi. Rozważone zostały trzy typy niepewności: wartości wejść i wyjść zdefiniowane w formie przedziałów, nakłady oraz efekty zdefiniowane na skali porządkowej oraz, dla problemów z modelem efektywności opartym na funkcjach wartości, dopuszczalne zakresy cząstkowych funkcji wartości. Wszystkie powyższe typy niepewności zostały uwzględnione zarówno w metodach opartych o programowanie matematyczne, jak i analizie stochastycznej.

W metodach dokładnych poszczególne typy niedokładności zostały uwzględnione w opisanym poniżej sposób.

Przedziałowe wartości nakładów i efektów. W problemach programowania matematycznego przedziałowe wartości wejść i wyjść zostały zastąpione wartościami dokładnymi reprezentującymi najbardziej lub najmniej korzystny scenariusz dla badanej jednostki w zależności od rozważanego typu wyników. Dla modeli identyfikujących najlepsze możliwe wyniki dla badanej jednostki (scenariusz optymistyczny) precyzyjne wartości wejść i wyjść dla badanej jednostki odpowiadają minimalnym wejściom i maksymalnym wyjściom z podanego przedziału. Dla pozostałych jednostek wartości dokładne równe są równo maksymalnym wartościom z danego przedziału dla wejść i minimalnym wartościom dla wyjść. Analogicznie, w przypadku modeli znajdujących najgorsze możliwe wartości dla danej jednostki, wartości dokładne reprezentują najmniej korzystny scena-

riusz, tj. największe nakłady i najmniejsze efekty, dla tej jednostki i najbardziej korzystny (najmniejsze nakłady i największe efekty) dla pozostałych jednostek.

Czynniki zdefiniowane na skali porządkowej. Dla czynników (nakładów lub efektów) zdefiniowanych na skali porządkowej znana jest jedynie kolejność jednostek na danym kryterium. W modelach matematycznych zapewnione zostało, że zdefiniowana kolejność zostaje zachowana. Aby uniknąć nieliniowości modelu, w modelu ilorazowym, wprowadzono do modelu dodatkowe zmienne reprezentujące iloczyn wagi danego czynnika porządkowego i jego wartości dla poszczególnych jednostek. Na tak zdefiniowanych zmiennych nałożono ograniczenia zapewniające ich monotoniczność zgodnie ze zdefiniowanym porządkiem. W modelu opartym na funkcjach wartości czynniki porządkowe uwzględnione zostały w podobny sposób, jednak w tym przypadku zmienna zastępcza nie reprezentuje iloczynu wagi i oceny jednostki, lecz iloczyn wagi danego czynnika i jego wartości cząstkowej funkcji. Dodatkowo, w modelu opartym o funkcje wartości, ograniczenia w modelu muszą uwzględniać różny kierunek monotoniczności dla nakładów i efektów oraz zapewniać, że wartość przypisana dla najgorszej jednostki jest dodatnia (większa od pewnej małej wartości ϵ). Dla najlepszej jednostki wartość przypisana do nowo utworzonej zmiennej nie może przekraczać wagi danego czynnika.

Dopuszczalne zakresy cząstkowych funkcji wartości. Dla modelu bazującego na funkcjach wartości rozważono również trzeci rodzaj niepewności: dopuszczalne zakresy cząstkowych funkcji wartości. Takie zakresy definiuje się za pomocą dwóch funkcji wartości reprezentujących górną i dolną granicę dopuszczalnego zakresu. Aby uwzględnić tego typu niepewność w modelach programowania matematycznego ponownie wprowadzono zmienne zastępcze, podobnie jak dla czynników porządkowych. Jeśli dany nakład lub efekt jest również zdefiniowany w formie przedziałów, w pierwszej kolejności wartości przedziałowe muszą zostać zastąpione precyzyjnymi w sposób opisany powyżej. Następnie, konieczne jest zapewnienie, że cząstkowa wartość dla każdej z jednostek mieści się w zdefiniowanym zakresie funkcji. Dodatkowo, wymuszony został monotoniczny kształt funkcji wartości, podobnie jak dla czynników porządkowych.

Metody oparte na symulacji Monte Carlo. Aby wyznaczyć indeksy akceptowalności dla metody DEA z nieprecyzyjnymi danymi, procedura próbkowania odbywa się w kilku etapach. W pierwszej kolejności uruchomiony zostaje algorytm Hit-And-Run, aby uzyskać określoną liczbę próbek wektorów wag, w taki sam sposób jak dla problemów z precyzyjnymi danymi. Kolejne etapy zależą od typu niepewności oraz modelu efektywności. W przypadku nakładów i efektów zdefiniowanych w formie przedziałów, niezależnie od modelu efektywności, generowane są próbki efektywności z przedziałów zdefiniowanych dla każdej z jednostek. Dla czynników porządkowym zastosowane zostało podejście SMAA-O. Bez utraty ogólności można założyć, że oceny jednostek dla czynników porządkowych zawarte są w przedziale $[0, 1]$. Z tego zakresu wylosowane zostaje K wartości (K – liczba jednostek). Następnie, wartości te zostają potraktowane jako precyzyjne oceny poszczególnych jednostek dla badanego czynnika, z uwzględnieniem podanej kolejności. W przypadku modelu opartego o funkcje wartości ze zdefiniowanym zakresem dopuszczalnych funkcji, w ostatnim kroku generowane są próbki cząstkowych wartości dla wszystkich jednostek. Mając dane precyzyjne oceny jednostek, wylosowane zostają wartości cząstkowych funkcji z przedziału pomiędzy górną a dolną funkcją ogra-

niczącą dopuszczalny zakres. Dodatkowo, zapewnione zostaje, że cząstkowe wartości wylosowane w każdej próbie zachowują monotoniczny porządek. Po wygenerowaniu próbek wag, ocen i wartości cząstkowych funkcji, wartość efektywności dla każdej jednostki obliczona zostaje zgodnie z wybranym modelem. Ostatecznie, wyznaczone zostają indeksy akceptowalności w oparciu o te wartości efektywności.

3 Badanie odporności dla problemów z hierarchiczną strukturą nakładów i efektów

W klasycznym podejściu, struktura nakładów i efektów rozważanych w metodzie DEA jest płaska. W niniejszej rozprawie rozważono również problemy, w których wejścia i wyjścia zorganizowane zostały w wielopoziomową, hierarchiczną strukturę. Takie podejście posiada kilka korzyści. Po pierwsze, istnieje możliwość łatwej modyfikacji i aktualizacji zbioru danych o nowe czynniki. Po drugie, taka struktura pozwala na dekompozycję problemów na mniejsze składniki, które są łatwiejsze do zarządzania, a wyniki bardziej precyzyjne. Po trzecie, w hierarchicznej strukturze istnieje możliwość modelowania interakcji nie tylko pomiędzy pojedynczymi czynnikami, ale także pomiędzy całymi kategoriami. Ograniczenia na wagi mogą zostać zdefiniowane na każdym poziomie hierarchii.

W ramach niniejszej rozprawy zaproponowano metody pozwalające na otrzymanie wyników badania odporności dla jednostek decyzyjnych, z użyciem modelu opartego na funkcjach wartości, dla każdej kategorii w hierarchii czynników. Ponadto, modele matematyczne pozwalające na wyznaczanie skrajnych odległości od najlepszej jednostki, pozycji w rankingach oraz weryfikacji występowania koniecznych i możliwych relacji preferencji zostały zaadaptowane tak, aby uwzględniać możliwość definicji ograniczeń na wagi na wszystkich poziomach hierarchii.

Dodatkowo, w rozprawie przedstawione zostały zależności pomiędzy odpornymi wynikami dla różnych kategorii. Poniżej przedstawiono przykładowe z nich:

- jeżeli dla wszystkich kategorii, będących bezpośrednimi dziećmi pewnej kategorii c , minimalna odległość do najlepszej jednostki jest zerowa, to dla kategorii c minimalna odległość do najlepszej jednostki również wynosi 0;
- Jeżeli najlepsza pozycja pewnej jednostki w rankingu efektywności dla wszystkich bezpośrednich podkategorii danej kategorii c jest pierwsza (lub ostatnia) to dla kategorii c najlepsza pozycja tej jednostki jest również pierwsza (lub ostatnia);
- Jeśli dla danej pary jednostek i kategorii c relacja koniecznej preferencji zachodzi w każdej podkategorii c to ta relacja musi zachodzić też w kategorii c .

W przypadku metod stochastycznych, estymacja indeksów akceptowalności odbywa się w sposób podobny do klasycznych problemów, jednak, wygenerowany wektor wag musi uwzględniać wagi wszystkich nakładów, efektów oraz kategorii na każdym poziomie rozważanej hierarchii. Dla każdej kategorii suma wag jej podkategorii musi być równa 1.

4 Metody badania odporności dla oceny efektywności z uwzględnieniem wielu scenariuszy analizy

W niektórych przypadkach ten sam zbiór jednostek decyzyjnych oceniany jest kilkakrotnie, biorąc pod uwagę różne punkty widzenia (scenariusze). Przykładowo, w niniejszej rozprawie rozważono efektywność lekarzy osobno dla grup pacjentów zgłaszających różne dolegliwości. W takim przypadku analizę odporności można przeprowadzić na dwóch poziomach. Z jednej strony wyniki badania odporności można przeanalizować dla każdego scenariusza osobno. Z drugiej strony, w niniejszej rozprawie zaproponowano miary pozwalające na agregację wyników uzyskanych dla poszczególnych scenariuszy. Wprowadzone miary opierają się na wyznaczeniu koniecznych i możliwych wyników, odpowiadających wynikom uzyskanym dla, odpowiednio, wszystkich i co najmniej jednego ze scenariuszy. Przykładowo, bazując na relacjach koniecznej preferencji uzyskanych dla poszczególnych scenariuszy można zweryfikować występowanie relacji koniecznie koniecznej preferencji występującej, gdy dla danej pary jednostek relacja koniecznej referencji występuje we wszystkich scenariuszach. Analogicznie, relacja możliwie koniecznej preferencji występuje, gdy dla danej pary jednostek konieczna preferencja występuje dla co najmniej jednego scenariusza. Podobne miary można zdefiniować na podstawie możliwej preferencji. W przypadku przedziałów pozycji w rankingu efektywności można wyznaczyć możliwie konieczny przedział pozycji jako przecięcie przedziałów pozycji, uzyskanych dla poszczególnych scenariuszy. Podobnie, możliwie możliwy przedział pozycji zdefiniowany został jako suma zakresów pozycji dla pojedynczych scenariuszy. Analogiczne miary można zaproponować dla skrajnych wartości efektywności czy odległości od najlepszej jednostki.

5 Wybór reprezentatywnego wektora wag w oparciu o wyniki badania odporności

W tradycyjnym podejściu metoda DEA wyznacza, dla każdej jednostki osobno, najbardziej korzystny, wektor wag. Zasadność porównywania jednostek w takim przypadku budzi wątpliwości z uwagi na brak wspólnej bazy do porównania. Z tego powodu, w niektórych zastosowaniach, korzystne jest znalezienie jednego, wspólnego, wektora wag dla wszystkich jednostek. W literaturze zaproponowano wiele takich metod.

W niniejszej rozprawie wprowadzona została nowa procedura wyznaczania wspólnego wektora wag w oparciu o wyniki badania odporności. Wynikiem zaproponowanej metody jest jeden, wspólny wektor wag, który możliwie najlepiej reprezentuje wszystkie dopuszczalne wektory wag. Konkretnie, jeśli wyniki badania odporności wskazują na wyraźną preferencję jednej jednostki nad inną, to różnica w wartościach miary efektywności dla tych jednostek powinna być możliwie największa. Z drugiej strony, dla par jednostek nieporównywalnych pod względem badanej miary, wartość efektywności po zastosowaniu uzyskanego wektora wag powinna być możliwie najmniejsza.

W tej rozprawie zaproponowano cztery różne relacje pozwalające na ocenę czy wyniki badania odporności wskazują na preferencję jednej jednostki nad drugą. Relacje te opierają się na, odpowiednio, oczekiwanej wartości efektywności, oczekiwanej pozycji w rankingu, relacji koniecznej preferencji oraz wartościach indeksu PEOI.

Oba, opisane powyżej, cele mogą zostać osiągnięte poprzez sekwencyjne rozwiązanie dwóch problemów programowania matematycznego reprezentujących poszczególne cele.

W pierwszej kolejności maksymalizowana jest najmniejsza z różnic dla par, dla których zidentyfikowano preferencję. Następnie, minimalizowana jest największa z różnic efektywności dla nieporównywalnych jednostek (z zachowaniem maksymalnej różnicy uzyskanej w poprzednim etapie). Wektor wag, uzyskany w optymalnym rozwiązaniu drugiego z modeli najlepiej reprezentuje całe spektrum dopuszczalnych wektorów wag.

6 Wyznaczanie reduktów i konstruktów efektywności

Analiza odporności dla metody DEA została również rozszerzona poprzez wprowadzenie dwóch nowych pojęć, które ułatwiają generowanie wyjaśnień wyników metody:

- redukt efektywności, dla pewnej efektywnej jednostki decyzyjnej, jest minimalnym zbiorem nakładów i efektów, które sprawiają, że dana jednostka jest efektywna;
- konstrukt efektywności, dla pewnej nieefektywnej jednostki decyzyjnej, jest najmniejszym możliwym zbiorem jednostek, które powodują jej nieefektywność. Inaczej mówiąc, jest to minimalny zbiór jednostek, które musiałyby zostać usunięte ze zbioru danych, aby badana jednostka została efektywna.

Do wyznaczenia reduktów efektywności zaproponowano przyrostowy algorytm, w którym sprawdzana jest efektywność badanej jednostki poczynając od najmniejszych (jednoelementowych) podzbiorów wejść i wyjść. W przypadku, gdy dla danego zbioru wejść i wyjść jednostka jest efektywna, wtedy wszystkie nadzbiory tego zbioru nie są brane pod uwagę w dalszym przeszukiwaniu. Dla każdej jednostki efektywnej można wyznaczyć co najmniej jeden redukt efektywności.

Dla jednostek nieefektywnych można wyznaczyć konstrukty efektywności. W tym celu rozwiązuje się ten sam model matematyczny, który pozwala na znalezienie minimalnej pozycji w rankingu efektywności. Jednostki, dla których zmienne binarne w optymalnym rozwiązaniu tego modelu przyjmują wartość 1 tworzą jeden z konstruktów efektywności. Kolejne konstrukty można znaleźć dodając do rozważanego modelu ograniczenie nie pozwalające na ponowne znalezienie tego samego rozwiązania.

7 Eksperymentalne porównanie metod tworzących pełen ranking jednostek decyzyjnych

Jak wspomniano wcześniej, klasyczne podejście w metodzie DEA pozwala jedynie na wyodrębnienie podzbioru jednostek efektywnych, bez możliwości porównania ich efektywności. Na przestrzeni ostatnich 50 lat, w literaturze zaproponowano liczne rozszerzenia pozwalające na konstrukcję pełnego rankingu jednostek decyzyjnych. W tej rozprawie przeprowadzono przegląd takich metod oraz ich eksperymentalne porównanie. Rozważone zostało 15 procedur pozwalających na uzyskanie rankingu jednostek decyzyjnych, spośród których 4 bazują na wynikach badania odporności i zostały oryginalnie zaproponowane w niniejszej pracy.

Metody tworzenia rankingu jednostek decyzyjnych wprowadzone w ramach tej pracy zostały oparte na wynikach metod badania odporności. Pierwsze dwie z nich porządkują zestaw jednostek w oparciu o malejące wartości oczekiwane miary efektywności oraz rosnące wartości oczekiwane pozycji w rankingu efektywności.

Pozostałe dwie miary tworzą ranking jednostek wykorzystując macierz indeksów PEOI. Procedura NFS-PEOI czerpie inspirację z miary Net FLow Score wykorzystywanej szeroko w MCDA m.in. w metodzie PROMETHEE. W tej metodzie, dla każdej jednostki, wyznaczany jest jej ogólny przepływ jako różnica przepływu dodatniego i ujemnego. Dodatni przepływ określa względną siłę danej jednostki (jej przewagę nad innymi jednostkami). Analogicznie, ujemny przepływ reprezentuje względną słabość badanej jednostki. Jednostki z najwyższą wartością ogólnego przepływu uznawane są za najlepsze.

Ostatnia metoda, wprowadzona w tej rozprawie, została nazwana PEV-PEOI i również opiera się na macierzy PEOI. W tej procedurze jednostki porządkowane są na podstawie wektora własnego odpowiadającego największej wartości własnej macierzy PEOI.

W niniejszej rozprawie wszystkie rozważone procedury porządkujące zostały opisane i zilustrowane na małym przykładzie. Następnie zidentyfikowano cechy poszczególnych metod i przedstawiono je w formie listy wad i zalet każdej z nich. Taka lista może pomóc w identyfikacji metody najbardziej pasującej do konkretnego problemu. Ostatecznie, rankingi wygenerowane przez poszczególne metody zostały porównane z wykorzystaniem pięciu różnych miar zgodności: współczynnika trafień (ang. Hit Ratio (HR)), znormalizowanego współczynnika trafień (ang. Normalized Hit Ratio (NHR)), τ Kendalla, miary różnic rankingów (ang. Rank Difference Measure (RDM)) oraz miary zgodności rankingów (ang. Rank Acceptance Measure (RAM)).

Eksperymenty składały się z dwóch części. W pierwszym etapie procedury zostały uruchomione dla 960 sztucznie wygenerowanych zbiorów danych z różną liczbą jednostek, nakładów i efektów. Następnie, uzyskane wyniki zostały porównane z tymi otrzymanymi dla 10 rzeczywistych zbiorów danych.

Wyniki uzyskane dla wszystkich pięciu miar były spójne, tj. identyfikowały te same pary metod jako generujących zbieżne rankingi. Spośród rozważanych procedur wyodrębniono trzy grupy metod, dla których generowane wyniki są podobne. Zidentyfikowano również trzy metody, które nie wpisują się w żadną z powyższych grup.

Z przeprowadzonej analizy wyraźnie wynika, że wybór metody wpływa, w istotny sposób, na otrzymany ranking. Wybierając procedurę generowania rankingów należy wziąć pod uwagę jej cechy oraz koncepcję, na której bazuje. Dodatkowo, korzystne wydaje się zestawienie wyników kilku metod, co pozwala na analizę problemu z różnych punktów widzenia.

8 Zastosowanie zaproponowanych metod do rzeczywistych problemów

Metodologiczna część niniejszej rozprawy zilustrowana została licznymi zastosowaniami dla rzeczywistych problemów.

Metody badanie odporności dla modelu ilorazowego zostały zaaplikowane do oceny efektywności 11 lotnisk w Polsce. Przedstawiono i opisano wyniki badania odporności w trzech różnych sytuacjach. W pierwszej kolejności rozważono najprostsza sytuację, w której brano pod uwagę wszystkie lotniska ze zbioru bez zdefiniowanych ograniczeń na wagi. Następnie wprowadzono dodatkowe ograniczenia na wagi. Na koniec zidentyfikowano i usunięto ze zbioru danych jednostki będące przykładami odstającymi. Analiza trzech powyższych sytuacji pokazała użyteczność proponowanych metod, wpływ defini-

cji ograniczeń wag na wyniki analizy oraz stosowalność metod badania odporności do wykrywania i usuwania jednostek będących przykładami odstającymi.

W kolejnej publikacji przedstawiono zastosowanie proponowanego podejścia do oceny niezawodności dostaw energii elektrycznej w 140 krajach. Problem został analizowany z kilku różnych perspektyw. Po pierwsze, zastosowano klasyczny ilorazowy model DEA do wskazania zbioru krajów efektywnych. W kolejnym etapie zastosowano zaproponowane algorytmy wyznaczania reduktów dla krajów efektywnych i konstruktów dla państw nieefektywnych. Ponadto, dla krajów nieefektywnych, wyznaczono jednostki będące referencjami (tzw. HCU) oraz poprawki konieczne do osiągnięcia efektywności. Następnie w pracy przedstawiono i omówiono wyniki metod badania odporności wprowadzonych w tej rozprawie. Na koniec wskazano i rozważono trzy potencjalne scenariusze rozwoju dla Japonii i Singapuru. Przeprowadzona analiza pokazała, że połączenie klasycznego modelu CCR z nowo zaproponowaną analizą odporności tworzy spójną całość i może zostać zastosowane w wielu dziedzinach.

Zastosowanie zaproponowanych metod z uwzględnieniem modelu opartego o funkcje wartości zostało zilustrowane na przykładzie analizy efektywności lekarzy z oddziału ratunkowego szpitala dziecięcego w Ottawie. Analiza została przeprowadzona osobno dla trzech grup pacjentów skarżących się na różne dolegliwości. W pracy przedstawiono zarówno szczegółowe wyniki badania odporności dla jednej z tych grup pacjentów jak i wyniki podejścia agregującego rezultaty uzyskane dla poszczególnych grup pacjentów. Dodatkowo, w ramach przeprowadzonej analizy, zastosowano algorytm pozwalający na wybór wektora wag reprezentującego wyniki badania odporności.

Kolejne trzy zastosowania przedstawione w niniejszej rozprawie dotyczyły analizy efektywności chińskich portów, robotów przemysłowych oraz Specjalnych Stref Ekonomicznych w Polsce. We wszystkich tych problemach uwzględniono nieprecyzyjne dane. W przypadku oceny portów i robotów przemysłowych zastosowano model ilorazowy, podczas gry do oceny Specjalnych Stref Ekonomicznych wykorzystano model oparty na funkcjach wartości.

Użyteczność zaproponowanych metod w kontekście hierarchicznej struktury nakładów i efektów zilustrowano dla problemu oceny systemów ochrony zdrowia w polskich województwach. Analiza została przeprowadzona 4 różnych poziomach: poprawa stanu zdrowia mieszkańców, efektywna gospodarka finansowa, satysfakcja pacjentów oraz całościowość systemu. W pracy pokazano sensowność analizy dla każdej z kategorii osobno oraz pokazano w jaki sposób agregacja takich wyników pozwala na identyfikację słabych i mocnych stron badanych jednostek.

Podsumowanie

Badania przedstawione w niniejszej rozprawie dotyczyły metod analizy odporności dla Granicznej Analizy Danych uwzględniających cały zakres dopuszczalnych wektorów wag nakładów i efektów.

Zaproponowane rozwiązanie składa się z dwóch, wzajemnie uzupełniających się grup metod. Pierwsza z nich pozwala na uzyskanie dokładnych wyników z użyciem programowania matematycznego. Druga polega na zastosowaniu symulacji Monte Carlo do oszacowania indeksów akceptowalności reprezentujących rozkłady badanych miar. W pierwszej grupie metod wyniki, chociaż dokładne, pokazują tylko wartości skrajne, występujące

rzadko. Metody stochastyczne dostarczają dodatkowej informacji o tym, jak wartości miar rozłożone są pomiędzy wartościami skrajnymi.

Efektywność jednostek decyzyjnych oceniona została z trzech różnych punktów widzenia. Pierwszy z nich dotyczy wartości miary efektywności i pozwala na wyznaczenie skrajnych wartości tej miary możliwych do osiągnięcia przez daną jednostkę, jej rozkładu i wartości oczekiwanej. Analogicznie, w perspektywie rankingów efektywności, zaproponowane podejście pozwala wyznaczyć skrajne pozycje poszczególnych jednostek, rozkład tych pozycji oraz ich wartości oczekiwane. Ostatnią rozważoną perspektywą były porównania parami jednostek. W tym przypadku zaproponowane modele pozwalają na ocenę występowania relacji koniecznej i możliwej preferencji dla par jednostek decyzyjnych przez prawdopodobieństwo wystąpienia preferencji jednej jednostki nad inną.

Rozważono dwa różne modele efektywności: ilorazowy, stosowany w klasycznym podejściu, oraz model addytywny oparty o funkcje wartości, w którym wartość efektywności wyznaczana jest jako ważona suma wartości cząstkowych funkcji zdefiniowanych dla poszczególnych nakładów i efektów. Ponadto, dla obu modeli efektywności wprowadzono rozszerzenia zaproponowanych metod uwzględniające problemy z nieprecyzyjnymi danymi. Rozważono trzy formy niepewności, tj. przedziały wartości nakładów i efektów, czynniki zdefiniowanych na skali porządkowej oraz dopuszczalne zakresy cząstkowych funkcji wartości. Modele matematyczne oraz metoda wyznaczania indeksów akceptowalności zostały zaadaptowane tak, aby uwzględniać wszystkie te typy niepewności.

Zaproponowane podejście pozwala również na analizę problemów, w których nakłady i efekty pogrupowane są w kategorii tworząc wielopoziomą strukturę hierarchiczną. W tym przypadku, metody badania odporności pozwalają na uzyskanie wyników dla każdej kategorii osobno. Takie rezultaty dają pozwalają na analizę efektywności działania jednostek z różnych perspektyw i precyzyjniejsze wyznaczenie obszarów, w których poszczególne jednostki działają dobrze oraz takich, które wymagają usprawnień. Dodatkowo, w niniejszej rozprawie zaproponowano zestaw miar pozwalających na dwuetapową analizę odporności przypadku, gdy ten sam zestaw jednostek oceniany jest dla różnych scenariuszy.

Aby uprościć analizę wyników badania odporności, które mogą być trudne do interpretacji zaproponowano metodę pozwalającą na wyznaczenie jednego zestawu wag nakładów i efektów w taki sposób, aby uzyskane wartości efektywności możliwie najlepiej reprezentowały odporne wyniki. W tym celu wprowadzono dwuetapową procedurę. W pierwszym etapie maksymalizowana jest różnica pomiędzy wartościami miary efektywności dla par jednostek, dla których zaobserwowano wyraźną preferencję. W drugim etapie minimalizuje się różnice efektywności dla par jednostek nieporównywalnych pod względem rozważanej miary.

Ostatnie rozszerzenie, wprowadzone w ramach tej pracy, ułatwia decydom wyjaśnianie wyników metody DEA. W tym celu zdefiniowano pojęcia reduktu i konstrukt efektywności i zaproponowano algorytmy pozwalające wyznaczyć redukt efektywności dla jednostek efektywnych i konstrukty dla jednostek nieefektywnych.

W niniejszej rozprawie zaproponowano cztery nowe procedury do tworzenia pełnego rankingu jednostek decyzyjnych oparte o wyniki badania odporności. Powyższe metody zostały zestawione i porównane z innymi procedurami, opisanymi w literaturze. Wyniki analizy wyraźnie pokazały, że wybór metody rankingowej wpływa w znaczny sposób na uzyskane rezultaty.

Zaproponowane metody badania odporności dla obu modeli efektywności zostały zaimplementowane i udostępnione w ramach otwarto-źródłowego projektu *diviz*. Kod źródłowy stworzonych modułów napisany jest w języku R i dostępny pod adresem https://github.com/alabijak/diviz_DEA/.

Niniejsza praca ma również charakter praktyczny. Wszystkie metody i rozszerzenia zostały zastosowane do rzeczywistych problemów dotyczących różnych dziedzin, m.in. ocena efektywności polskich lotnisk, niezawodności dostaw energii elektrycznej krajów oraz pracy lekarzy z oddziału ratunkowego szpitala dziecięcego w Ottawie.

Declarations

June 2, 2023

Anna Labijak-Kowalska
Institute of Computing Science
Poznań University Of Technology
Piotrowo 2
60-965 Poznań

Declaration

I hereby declare the following contribution as an author of the papers:

P. Gasser, M. Cinelli, A. Labijak, M. Spada, P. Burgherr, M. Kadziński, and B. Stojadinović. Quantifying electricity supply resilience of countries with robust efficiency analysis. *Energies*, 13(7):1535, 2020, DOI: 10.3390/en13071535

- Co-authorship of the idea of application of robustness analysis methods for DEA in the study of analysis of the resilience of countries' electricity systems,
- application of the ratio-based DEA to obtain the sets of efficient and inefficient countries, projections onto the efficient frontier, and necessary improvements for inefficient countries,
- application of the robustness analysis methods to the data considered in the study,
- implementation of the software necessary in the study,
- authorship of the concept and algorithms for determination of efficiency reducts and costructs,
- preparation of the updated results for the three scenarios for countries development,
- consultations on the data preparation and interpretation of the results,
- reviewing and correcting the text of the manuscript in terms of the methodology of the analysis.

A. Labijak-Kowalska, M. Kadziński, I. Sychała, L. C. Dias, J. Fiallos, J. Patrick, W. Michalowski, and K. Farion. Performance evaluation of emergency department physicians using robust value-based additive efficiency model. *International Transactions in Operational Research*, 30(1):503–544, 2023:

- co-authorship of the idea of adaptation of robustness analysis framework for the value-based DEA model,
- authorship of the mathematical models for the robustness analysis for VDEA model,
- authorship of the procedure for estimation the acceptability indices using Hit-And-Run algorithm in a context of value-based DEA model,
- authorship of the proposed measures for multiple scenarios of efficiency analysis,
- implementation of the software necessary in the study,
- application of the robustness analysis methods for the case study,
- analysis and discussion of the results,
- authorship of the first draft of the manuscript.


Anna Labijak-Kowalska

Poznań, 15 października 2016r.

Małgorzata Napieraj
JCommerce SA
ul. Rataje 164
61-168 Poznań

OŚWIADCZENIE DOTYCZĄCE WKŁADU W POWSTANIE PRACY

M. Kadziński, A. Labijak, M. Napieraj, Integrated framework for robustness analysis using ratio-based efficiency model with application to evaluation of Polish airports.
Omega, doi:10.1016/j.omega.2016.03.003, 2016

- implementacja wybranych modułów analizy odporności dla granicznej analizy danych dedykowanych na platformę diviz (Rozdział 3.2);
- przygotowanie w postaci tabelarycznej wybranych wyników analizy odporności dla problemu analizy efektywności polskich lotnisk (Rozdział 4.3);
- szacowany wkład procentowy: 5%.

Małgorzata Napieraj
Małgorzata Napieraj

June 2, 2023

Patrick Gasser
AFRY
Herostrasse 12
8048 Zürich
Switzerland

Declaration

I hereby declare the following contribution as an author of the paper:

P. Gasser, M. Cinelli, A. Labijak, M. Spada, P. Burgherr, M. Kadziński, and B. Stojadinović. Quantifying electricity supply resilience of countries with robust efficiency analysis. *Energies*, 13(7):1535, 2020, DOI: 10.3390/en13071535

- Collection of the data for the considered case study;
- Selection of the indicators for the case study;
- Co-authorship of the text of the publication in the application-oriented parts;
- Interpretation and discussion of the results.

A handwritten signature in blue ink that reads "PGasser". The letters are cursive and somewhat stylized.

Patrick Gasser

June 1, 2023

Marco Cinelli

Faculty Governance and Global Affairs
Leiden University College
Leiden, The Netherlands

Declaration

I hereby declare the following contribution as an author of the paper:

P. Gasser, M. Cinelli, A. Labijak, M. Spada, P. Burgherr, M. Kadziński, and B. Stojadinović. Quantifying electricity supply resilience of countries with robust efficiency analysis. *Energies*, 13(7):1535, 2020, DOI: 10.3390/en13071535

- Co-authorship of the idea underlying the paper consisting of applying the robust efficiency method to the considered study,
- Collection and preparation of the data for the considered study,
- Interpretation and discussion of the results,
- Reviewing and corrections on the text of the publication.



Marco Cinelli, PhD

June 2, 2023

Matteo Spada

Zurich University of Applied Sciences (ZHAW)
School of Engineering
Institute of Sustainable Development
Technoparkstrasse 2
8400 Winterthur

Declaration

I hereby declare the following contribution as an author of the paper:

P. Gasser, M. Cinelli, A. Labijak, M. Spada, P. Burgherr, M. Kadziński, and B. Stojadinović. Quantifying electricity supply resilience of countries with robust efficiency analysis. *Energies*, 13(7):1535, 2020, DOI: 10.3390/en13071535

- Co-authorship of the idea underlying the paper consisting of analyzing electricity supply resilience of countries;
- Supervision of the data collection and indicator selection process for the considered study;
- Reviewing and correcting the text of the manuscript.

Matteo Spada



June 2, 2023

Dr. Peter Burgherr

Head Technology Assessment Group
Laboratory for Energy Systems Analysis
Paul Scherrer Institut (PSI)
Forschungsstrasse 111
OHSA D18
5232 Villigen PSI
Switzerland
peter.burgherr@psi.ch

Declaration

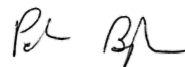
I hereby declare the following contribution as an author of the paper:

P. Gasser, M. Cinelli, A. Labijak, M. Spada, P. Burgherr, M. Kadziński, and B. Stojadinović. Quantifying electricity supply resilience of countries with robust efficiency analysis. *Energies*, 13(7):1535, 2020, DOI: 10.3390/en13071535

- Overall coordination as responsible Principal Investigator within the Future Resilient Systems (FRS) program of the Singapore ETH Centre (SEC).
- Co-authorship of the idea underlying the paper consisting of analyzing electricity supply resilience of countries;
- Supervision of the data collection and indicator selection process for the considered study;
- Reviewing and correcting the text of the manuscript.

Villigen PSI, 02.06.2023

Place, Date



Dr. Peter Burgherr
Head Technology Assessment Group

May 31, 2023

Inga Spychała
Instytut Informatyki
Politechnika Poznańska
Piotrowo 2
60-965 Poznań

Declaration

I hereby declare the following contribution as an author of the paper:

Labijak-Kowalska, M. Kadziński, I. Spychała, L. C. Dias, J. Fiallos, J. Patrick, W. Michalowski, and K. Farion. Performance evaluation of emergency department physicians using robust value-based additive efficiency model. *International Transactions in Operational Research*, 30(1):503–544, 2023:

- Co-authorship of the models for finding the representative set of weights based on the outcomes of the robustness analysis,
- Implementation of the procedures for finding the common set of weights.


Inga Spychała

May 31, 2023

Luis C. Dias
University of Coimbra
CeBER, Faculty of Economics
Coimbra, Portugal

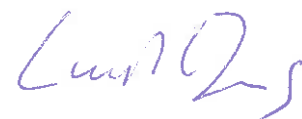
Declaration

I hereby declare the following contribution as a co-author of the following papers:

A. Labijak-Kowalska, M. Kadziński, I. Spychała, L. C. Dias, J. Fiallos, J. Patrick, W. Michalowski, and K. Farion. Performance evaluation of emergency department physicians using robust value-based additive efficiency model. *International Transactions in Operational Research*, 30(1):503–544, 2023.

A. Labijak-Kowalska, M. Kadziński, and L. C. Dias. Robustness analysis for imprecise additive value efficiency analysis with an application to evaluation of special economic zones in Poland, submitted to *Socio-Economic Planning Sciences*, 2023.

- Co-authorship of the idea underlying the papers consisting in conducting robustness analysis in the context of value-based efficiency analysis;
- Consultations on the proposed efficiency-based concepts and respective mathematical models (I was previously a co-author of the value-based efficiency model that serves as the basis for the extensions presented in these papers);
- Review and editing of the manuscripts.



Luis C. Dias

May 31, 2023

Javier Fiallos
Elizabeth Bruyere Hospital
43 Bruyere St., Ottawa, ON,
K1N 5C8 Canada

Declaration

I hereby declare the following contribution as an author of the following paper:

A. Labijak-Kowalska, M. Kadziński, I. Spychała, L. C. Dias, J. Fiallos, J. Patrick, W. Michalowski, and K. Farion. Performance evaluation of emergency department physicians using robust value-based additive efficiency model. *International Transactions in Operational Research*, 30(1):503–544, 2023.

- Collection of the data used in the study of evaluating the performance of emergency department physicians;
- Consulting the results obtained in the study;
- Review and editing of the manuscript.



Javier Fiallos

May 31, 2023

Wojtek Michalowski

Telfer School of Management
University of Ottawa
55 Laurier Ave. E, Ottawa, ON
K1N 6N5 Canada

Declaration

I hereby declare the following contribution as an author of the following paper:

A. Labijak-Kowalska, M. Kadziński, I. Spychała, L. C. Dias, J. Fiallos, J. Patrick, W. Michalowski, and K. Farion. Performance evaluation of emergency department physicians using robust value-based additive efficiency model. *International Transactions in Operational Research*, 30(1):503–544, 2023.

- Supervision of collecting data used in the study of evaluating the performance of emergency department physicians;
- Co-authorship of the idea of using Data Envelopment Analysis in the context of the case study;
- Consulting the results obtained in the study;
- Review and editing of the manuscript.



Wojtek Michalowski

May 31, 2023

Weronika Mrozek
Instytut Informatyki
Politechnika Poznańska
Piotrowo 2
60-965 Poznań

Declaration

I hereby declare the following contribution as an author of the paper:

Labijak-Kowalska, M. Kadziński, and W. Mrozek. Robust additive value-based efficiency analysis with a hierarchical structure of inputs and outputs. *Applied Sciences*, 13(11), 2023,

DOI: 10.3390/app13116406

- Co-authorship of the idea underlying the paper,
- Reviewing and editing the text of the publication,
- Preparation of part of the software for evaluating the healthcare systems in Poland.
- Visualization of some results of the case study.

Mrozek

Weronika Mrozek



© 2023 Anna Labijak-Kowalska

Poznan University of Technology
Faculty of Computing and Telecommunications
Institute of Computing Science
Typeset using L^AT_EX in Computer Modern.

Bib_TE_X:

```
@phdthesis{ Labijak-Kowalska2023,  
  author = "Anna Labijak-Kowalska",  
  title = "New Directions  
in Robust Data Envelopment Analysis",  
  school = "Poznan University of Technology",  
  address = "Pozna{\n}, Poland",  
  year = "2023",  
}
```