Expert Systems With Applications 296 (2026) 129010

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Enhancing the Benefit of the Doubt model through 'Ensemble-DEA': achieving the Sustainable Development Goals

Juan Aparicio [a,b], Magdalena Kapelko [c,*], Juan F. Monge [a], José L. Zofío [d,e]

[a] Center of Operations Research, Miguel Hernandez University of Elche, Avda. de la Universidad s/n, E-03202 Elche, Spain
[b] valgrAI – Valencian Graduate School and Research Network of Artificial Intelligence, Valencia, Spain
[c] Department of Logistics, Wroclaw University of Economics and Business, Wroclaw, Poland
[d] Department of Economics, Universidad Autónoma de Madrid, Madrid, Spain
[e] Erasmus Research Institute of Management, Erasmus University, Rotterdam, the Netherlands

ARTICLE INFO

Keywords:
Benefit of the Doubt (BoD)
Data Envelopment Analysis (DEA)
Ensemble of models
Randomization
Composite indicators

ABSTRACT

This study presents an innovative approach for constructing composite indicators by combining the Benefit of the Doubt method from Data Envelopment Analysis with ensemble techniques, i.e., 'Ensemble-DEA', with randomization in observations and variables selection. Our methodology mitigates the curse of dimensionality, which limits the effectiveness of traditional approaches when dealing with numerous indicators. By maintaining data integrity and improving robustness through an ensemble-based technique, our method delivers high-discriminatory power and clear rankings for Decision Making Units. Additionally, it enhances benchmarking capabilities by offering unit-specific peer comparisons. Our contributions therefore include the development of robust composite indicators and improved benchmarking insights, ensuring their reliability even in high-dimensional settings. We validate our approach using a real-world dataset containing 72 indicators aligned with Sustainable Development Goals for European Union countries. The results show that performance in meeting Sustainable Development Goals is correlated with the level of socioeconomic development and environmental consciousness. In particular, Scandinavian, Northern European and Benelux countries tend to perform best, while Eastern European countries lag in sustainability effectiveness. Furthermore, a comparative analysis against conventional methods underscores the advantages of our approach in managing complex datasets, specifically in terms of improvement in discriminatory power and benchmarking opportunities.

1. Introduction

Data Envelopment Analysis (DEA) is a non-parametric methodology used to evaluate the relative efficiency of Decision Making Units (DMUs) that utilize multiple inputs to produce multiple outputs. Originally developed by Charnes et al. (1978) and Banker et al. (1984), DEA has become a fundamental tool in efficiency analysis, enabling the direct comparison of homogeneous units. By constructing a piecewise linear frontier over the data points, DEA identifies efficient units that form the efficient frontier, against which the performance of other units is measured. This method not only highlights inefficiencies but also provides specific targets and reference sets for improvement, fostering a deeper understanding of operational performance. DEA's benchmarking capabilities are particularly valuable in today's dynamic and performance-driven environments.

Despite the advantages of DEA, distinguishing between efficient and inefficient DMUs can become problematic when there are many variables relative to the number of DMUs. This imbalance can result in inefficient units being mistakenly identified as efficient ones. The limited discriminative power due to this issue, that is, the impossibility of ranking DMUs according to their efficiency score in a suitable manner, has significant consequences, as it can restrict the practical managerial insights that can be obtained (Ghasemi et al., 2019). Overall, the higher the number of variables relative to the number of observations, the less effective the DEA analysis becomes in distinguishing efficiency levels (Jenkins & Anderson, 2003). The reason is that DEA methods search for the most favourable solution for the unit under evaluation and increasing the dimensionality of the analysis offers a larger chance of adjusting the weights when evaluating the relative efficiency. This problem, known in the literature as the curse of

* Corresponding author.
E-mail addresses: j.aparicio@umh.es (J. Aparicio), magdalena.kapelko@ue.wroc.pl (M. Kapelko), monge@umh.es (J.F. Monge), jose.zofio@uam.es (J.L. Zofío).

https://doi.org/10.1016/j.eswa.2025.129010
Received 7 February 2025; Received in revised form 30 June 2025; Accepted 11 July 2025
Available online 12 July 2025

dimensionality, can ultimately result, as we illustrate in our empirical application, in all DMUs being efficient.

As for this issue, the literature indicates some 'heuristic' (or empirical) rules regarding the number of DMUs versus the number of inputs and outputs to solve the above problem. Specifically, the guideline from Homburg (2001) and Golany & Roll (1989) suggests that the number of DMUs should be at least twice the number of variables. Similarly, research by Nunamaker (1985), Raab and Lichty (2002), Friedman & Sinuany-Stern (1998), and Banker et al. (1989) recommends that the number of observations should be at least three times the number of variables. On the other hand, Dyson et al. (2001) pointed out that the number of units should be at least twice the product of the number of inputs and outputs. Additionally, Cooper et al. (2007) proposed another 'rule of thumb': $n \geqslant \max\{m \cdot s, 3 \cdot (m+s)\}$, where n is the sample size, m is the number of inputs and s is the number of outputs. These heuristics help improve the discrimination power of DEA scores. However, in most practical instances, the dataset is predefined, and the analyst cannot modify it to comply with these rules.

A second way to mitigate the curse of dimensionality and its consequences consists in reducing the number of variables to be considered in the analysis. In the extensive body of literature on variable selection, numerous methodologies have been proposed (for a thorough review, see Nataraja & Johnson, 2011). Here, we highlight some of the more widely adopted techniques. The Efficiency Contribution Measure (ECM) evaluates the significance of variables by analyzing their relative contribution to efficiency scores, as discussed by Pastor et al. (2002). Another notable approach is the Regression-Based (RB) method introduced by Ruggiero (2005), which augments a basic model by incorporating variables that exhibit a positive correlation with the efficiency scores derived from the initial model. Further advancements in this field include methods that extend traditional DEA linear programming models by incorporating binary variables to determine variable inclusion or exclusion. Examples of such methodologies are found in the works of Limleamthong & Guillén-Gosálbez (2018), Peyrache et al. (2020), and Benítez-Peña et al. (2020). Moreover, other approaches enhance corrected Concave Nonparametric Least Squares (C2NLS) (Kuosmanen & Johnson, 2010) with LASSO-like regularization terms as demonstrated by Lee and Cai (2020) and Chen et al. (2021). Also, the method proposed by Wilson (2018) based on the usage of eigenvalue decomposition of the moment matrices of inputs and outputs is another popular proposal for dimension reduction. Additionally, some studies have explored the integration of Principal Component Analysis (PCA) with DEA, such as the approach described by Adler & Golany (2001).

Additionally, a third alternative to get a suitable ranking of units, that is, increasing the discrimination power of the efficiency scores, consists in modifying the standard DEA model, maintaining the original structure of variables and data. In this regard, a comprehensive review of this type of literature by Aldamak & Zolfaghari (2017) classifies existing methods into ten distinct categories. The first category includes Cross-Efficiency methods (Doyle & Green, 1994), while the second pertains to the Superefficiency approach introduced by Andersen and Petersen (1993). The third category ranks units based on their relative importance to inefficient units (Torgersen et al., 1996). The fourth category applies statistical techniques, like canonical correlation analysis, directly after running a DEA model (Friedman & Sinuany-Stern, 1997). The fifth category concentrates on ranking inefficient DMUs using a 'measure of efficiency dominance' as suggested by Bardhan et al. (1996). The sixth category employs multilevel DEA within multicriteria decision-making methods, supplemented with analytic hierarchy processes, as described by Jablonsky (2007). The seventh category incorporates inefficient frontiers by solving 'inverted' DEA models (Yamada, 1994). In the eighth category, the TOPSIS approach considers virtual units that identify the best (ideal) and worst (anti-ideal) performance (Barbero et al., 2021). The ninth category involves methods where decision-makers incorporate external judgments into the evaluation process through weight restrictions (see, for example, Allen et al., 1997). The tenth category involves fuzzy methods (Angiz et al., 2013). Among all these methods, the most used in practice are those based on Superefficiency and Cross-Efficiency, due to their foundational role and enduring relevance in performance analysis.

A relevant application of DEA is in the construction of composite indicators through the method known as Benefit of the Doubt (BoD) (Cherchye et al., 2004, 2007a, 2007b). Composite indicators synthesize multidimensional data into a single index, providing a comprehensive measure of performance. BoD approach has been used in numerous settings as it allows an objective (unsupervised) aggregation of multiple indicators of different nature, thereby enhancing their reliability and interpretability. It provides the DEA flexibility in weighting the different dimensions according to their relative importance when measuring performance, ensuring that the resulting composite indicator reflects the true performance and priorities of the evaluated units. Moreover, the construction of composite indicators is crucial for decision-making because it simplifies complex data.

From a methodological standpoint, the BoD approach measures the effectiveness of observations when reaching a stated goal.[1] Numerous advancements have been made to the original BoD model in DEA. Enhancements include models that incorporate slacks (Sahoo & Acharya, 2012), that include imprecise data (Cherchye et al., 2011) and those that account for preference structures among indicators (Fusco, 2015). Other notable models are those that include undesirable indicators (Zanella et al., 2015), robust conditional models that consider external factors (De Witte & Rogge, 2010, 2011), and non-compensatory composite indicators (Vidoli et al., 2015). Further developments include the spatial robust BoD model (Fusco et al., 2018) and intra- and inter-group BoD models (Karagiannis & Karagiannis, 2018). The inter-group BoD model based on the 'Global Frontier Difference Index' (Van Puyenbroeck & Rogge, 2020), the translation-invariant directional distance function model (Aparicio & Kapelko, 2019), and models integrating goal programming (Oliveira et al., 2019) are also significant. Additionally, there are models aimed at minimizing distance to the frontier (Aparicio et al., 2020) and methods incorporating spatial dependence into robust directional models for undesirable outputs (Fusco et al., 2020). Recent research has introduced models based on multiple attribute utility theory (Lahouel et al., 2021), integrating regulatory constraints, undesirable outputs and imperfect knowledge of data (Ferreira et al., 2023), and the integration of BoD with multi-directional efficiency analysis (Fusco, 2023). The most recent BoD developments include multidirectional robust BoD (Vidoli et al., 2024), sequential BoD (Walheer, 2024), common weights multiplicative approach (Koronakos et al., 2024) and the nonconvex BoD in multiplier form for comparison across groups of units (Kapelko et al., 2024). In addition, studies have explored aggregating DMU-level BoD scores to group-level scores, with contributions by Karagiannis (2017) and Rogge (2018a, 2018b).

Ultimately, composite indicators provide a clear, aggregated measure that can inform policy, drive improvements, and foster accountability. However, as the number of indicators is relatively large with respect to the number of observations (e.g., countries), BoD analysis suffers from the weakness related to the curse of dimensionality. This abundance of indicators can lead to challenges in effectively analyzing and comparing the data. For example, in scenarios where the number of indicators significantly exceeds the number of DMUs, traditional analytical methods may struggle to provide clear insights. These scenarios hinder the ability to discern true performance differences among

[1] Prieto & Zofío (2001) are among the first authors to apply the BoD method to evaluate effectiveness in the public provision of infrastructure and equipment by Spanish municipalities. They argue that models that do not account for the resources (inputs) employed to reach a goal cannot represent efficiency, as they only evaluate to what extent goals are met, but do not consider the relative efforts made to reach them. Despite this qualification, the term efficiency is still commonly used in BoD studies.

the DMUs.

In this study, we show that in these scenarios even widely used techniques to enhance the discriminatory power of BoD and DEA scores do not perform well. This is due to a combination of several factors: the usual interpretation of efficiency scores is lost, benchmarking information—which is crucial to improve public policies—cannot be determined, and/or the discriminatory power between units remains relatively low. These limitations underscore the necessity for more robust approaches to handle complex datasets with numerous indicators. To address this gap, we propose an innovative method, termed 'Ensemble-DEA', that combines systematic resampling of both observations and variables. Considering the empirical illustration of our method, this dual-randomization approach involves constructing an ensemble of BoD models, each estimated on a randomly selected subset of countries and performance indicators. Hence, the method can be regarded as a crossover between model ensemble and randomization with respect to DMUs and indicators, bridging the concepts often employed in machine learning and data science. Unlike conventional bootstrapping in DEA—which only samples over observations and suffers from performance collapse due to the curse of dimensionality—our method enhances discrimination by simultaneously reducing the effective dimensionality of each model while preserving the information diversity through repeated sampling. Regarding bootstrapping in the field of non-parametric efficiency measurement, it is worth noting that this technique has revolutionized the literature by offering a powerful tool for inference in DEA (see e.g., Simar & Wilson, 2000). Despite its utility and widespread adoption, bootstrapped DEA encounters significant limitations when applied to scenarios involving the construction of composite indicators with a high number of variables relative to the number of units being evaluated. In such contexts, the sub-samples generated through the Bootstrapping method may include a vast number of variables, which, due to the curse of dimensionality, can result in efficiency scores of one for all DMUs across all subsamples. This lack of variability in efficiency scores precludes the accurate correction of bias and the calculation of reliable confidence intervals, fundamentally undermining the efficacy of bootstrapped DEA in these settings (as we show in the empirical section). Our research addresses this critical gap by proposing an innovative solution through the ensemble of multiple BoD models, each of them using different random subsets of data and indicators. This strategy ensures that the resulting composite scores are interpretable in standard BoD terms and that meaningful benchmarking relationships can still be derived across the ensemble.

This study makes several contributions to the literature on composite indicators, particularly those based on the BoD approach. First, we extend the traditional BoD model by developing a novel ensemble-based framework within Data Envelopment Analysis, 'Ensemble-DEA', which combines multiple BoD estimations through a process of systematic resampling—simultaneously across observations and indicators. Second, our method addresses the critical issue of low discrimination power among DMUs, offering a more effective evaluation mechanism. Third, our approach generates scores that retain their efficiency or effectiveness interpretation in conventional terms, facilitating a clearer understanding and practical application of the results. Fourth, by providing detailed benchmarking information, including the identification of peers for each DMU evaluated, our method empowers organizations with important insights to drive performance improvements. Fifth, a notable contribution is the robustness of our composite indicator, which remains reliable even when incorporating a large number of variables, a common challenge in complex datasets. Our method achieves this robustness by randomly selecting subsets of indicators from the complete set. By repeating this resampling process multiple times with different randomly selected subsets of the same size, we ensure that the final composite indicator is not overly dependent on any specific set of variables. Sixth, our approach also introduces a new alternative for improving the ranking of units through increasing the discriminatory power of the approach. In particular, we offer a novel solution that has not been previously applied by incorporating ensemble methods and a process of double randomization (in data and indicators). Seventh, we undertake comprehensive comparative analysis of our method with the most popular methods from the literature, not only underscoring the advantages of our proposed method but also providing a thorough evaluation against established techniques. Finally, we demonstrate the practical applicability of our method with a real-world case study using data from the Sustainable Development Goals (EUROSTAT, 2023). This case study illustrates the efficacy and relevance of our approach in addressing contemporary challenges in public policy and performance evaluation.

The remainder of this paper is structured as follows. In Section 2, we provide the background necessary to understand our contributions. Section 3 introduces our novel approach. In Section 4, we apply our method to a dataset from EUROSTAT (2023), composing an indicator from 72 variables related to the Sustainable Development Goals and measuring the effectiveness of 27 countries when reaching this goal. Finally, Section 5 concludes the paper and outlines study implications and several potential future research directions.

2. Background

To establish a solid foundation for our contributions, we delve into the pertinent theoretical and methodological frameworks that underpin our research. This section begins by exploring the BoD approach based on data Envelopment Analysis and provides an overview of ensemble techniques with a particular focus on double randomization of the data: observations and variables.

2.1. The BoD approach to build composite indicators

The BoD approach (Cherchye et al., 2004, 2007a, 2007b; Lovell & Pastor, 1999; Prieto & Zofío, 2001) has emerged as a reliable methodology for constructing composite indicators, particularly in contexts where multiple dimensions of performance (indicators) need to be assessed. BoD's primary advantage lies in its data-driven weighting scheme. Rather than imposing a uniform set of weights across all units, BoD allows each unit to determine the most advantageous weights within a constrained optimization framework. This approach mitigates biases that can arise from arbitrary, subjective or expert-imposed weights, ensuring that the composite indicator truly represents the underlying data's multidimensional nature.

The BoD approach represents a specific subset of DEA methods, corresponding to a performance model that does not require inputs, treating all indicators as outputs to create a comprehensive and objective aggregated final indicator for each unit evaluated. Let us assume that n DMUs must be evaluated. DMU_i, $i = 1, \ldots, n$, has s associated performance indicators $\boldsymbol{y}_i = \left(y_{i1}, \ldots, y_{is}\right) \in R^s_+$.[2] Resorting to the BoD approach, the radial (effectiveness) score of DMU k is determined by the following linear optimization model (Lovell & Pastor, 1999)[3]:

[2] Non-bold is used for denoting scalars and bold is utilized for denoting vectors.

[3] With model (1) we follow Lovel & Pastor (1999) and Prieto & Zofío (2001) to estimate an output-oriented variable returns to scale (VRS) model without inputs. The BoD model put forward by Cherchye et al. (2004) is formulated as an input-oriented constant returns to scale (CRS) DEA model in the multiplier form with a common single constant input with a value at one. Van Puyenbroeck (2018) remarks that there is a direct, reciprocal relation between both formulations.

$$\begin{aligned} \phi_k^* = \quad & \max \quad && \phi_k \\ & s.t. \\ & && \sum_{i=1}^{n} \lambda_{ki} y_{ir} \geqslant \phi_k y_{kr}, \quad r = 1, \ldots, s, \\ & && \sum_{i=1}^{n} \lambda_{ki} = 1, \\ & && \lambda_{ki} \geqslant 0, \qquad i = 1, \ldots, n. \end{aligned} \tag{1}$$

The optimal solution ϕ_k^* is always greater than or equal to one, with $\phi_k^* = 1$ indicating the best performance (efficient DMU under the typical DEA terminology).[4] Additionally, the optimal solutions of model (1) can be used to get benchmarking information. Identifying peers involves analyzing the intensity variables, which play a crucial role in the envelopment process. These variables, corresponding to the optimal lambdas (λ_{ki}^*) obtained when solving (1), are coefficients that indicate the proportionate contribution of each efficient DMU in forming a reference point, either for itself, with a value of one, λ_{kk}^*=1, or for an inefficient DMU, $\lambda_{ki}^* > 0$. When a DMU is evaluated, the BoD model generates this set of intensity variables that define a linear combination of efficient units, which serve as benchmarks or peers. These intensity variables provide valuable insights into how an inefficient DMU can improve. For instance, if a particular peer has a high intensity variable (i.e., the closest it is to one), it suggests that this peer's practices are highly relevant and potentially beneficial for the underperforming unit to emulate.

2.2. Ensemble techniques

Ensemble techniques represent a fundamental paradigm in modern statistics, demonstrating a significant capacity to enhance predictive accuracy. These methods strategically combine the results of multiple base learners, capitalizing on the principle that a collection of diverse models can compensate for individual weaknesses, leading to more reliable and stable predictions. This improvement is often attributed to a reduction in both variance and bias, while effectively mitigating overfitting and enhancing the model's ability to generalize to unseen data (see Hastie et al., 2009).

Within the broader class of resampling-based methodologies, techniques that incorporate randomization over both units and variables have shown strong potential for improving model robustness and generalizability. In particular, strategies that generate multiple estimations by sampling different subsets of the original dataset—both in terms of the observations and the dimensions (indicators)—allow for the construction of ensembles that capture diverse structural relationships within the data. By aggregating the results across these repeated estimations, such approaches mitigate the limitations associated with high-dimensional settings and reduce the influence of any specific data configuration. This dual-randomization mechanism is especially effective in settings where maintaining the full complexity of the dataset is essential, yet standard methods fail to discriminate meaningfully across units due to dimensionality constraints.

Finally, and regarding the research gap addressed in this paper, it is worth mentioning that while DEA and BoD are powerful tools for efficiency analysis and composite indicator construction, they face significant challenges when dealing with high-dimensional datasets due to the curse of dimensionality. This issue arises when the number of variables is large relative to the number of observations, leading to a loss of discriminatory power and making it difficult to effectively rank DMUs. Existing methods to address this issue often involve reducing the number of variables, which can lead to a loss of information and potentially distort the results by excluding relevant dimensions. Other methods, while preserving the original data, result in efficiency scores that lack clear interpretation in terms of potential efficiency improvements or fail to provide useful benchmarking information for policy implications. To address these limitations and provide a more robust and interpretable approach, our research proposes a novel methodology that integrates the BoD approach with ensemble techniques to effectively handle high-dimensional data, while maintaining the interpretability of the results and providing useful benchmarking information.

3. The new approach to build BoD composite indicators based on 'Ensemble-DEA'

Building on the foundational concepts of the DEA methods applied to the BoD approach and the power of ensemble methods, we present a novel methodology for constructing composite indicators. This innovative approach exploits the flexibility and data-driven weighting of DEA with the predictive robustness and accuracy of ensemble models, creating a sophisticated framework for multidimensional performance assessment. By integrating DEA and ensemble techniques, our method not only enhances the reliability and validity of composite indicators but also offers deeper insights of the assessment problem by providing benchmarking information and scores with high discriminatory power among units. This section delves into the methodological underpinnings and practical implementation of our approach, illustrating how it can be effectively applied to datasets.

The essence of the new approach lies in combining multiple estimations of the BoD model through a structured process of double randomization—applied simultaneously to observations and indicators—which enhances robustness and discriminatory power in high-dimensional settings. This strategy is conceptually related to ensemble procedures that construct multiple models based on systematically varied subsets of the data. In our application, each BoD model is built on a different random subset of countries and indicators, and the final performance measure is obtained by aggregating the results. Although this resembles certain ensemble ideas used in other domains—such as the use of diverse model replications to stabilize results—our method remains entirely within the non-parametric DEA framework, without relying on predictive modelling or algorithmic training. This ensemble-based BoD structure, which we refer to as 'Ensemble-DEA', enables meaningful efficiency evaluation and composite indicator construction even in high-dimensional environments.

Before presenting the specifics of the proposed approach, we recall the discussion in the Introduction, positioning our method within the broader context of existing strategies designed to tackle the challenges posed by the curse of dimensionality, a well-known issue in DEA that hampers the ability to rank DMUs effectively. While some methods address this problem by reducing the number of variables or altering the dataset structure, others, such as Superefficiency or Cross-Efficiency models, attempt to resolve it without modifying the original data. However, our approach introduces a fundamentally new avenue by applying ensemble techniques, specifically tailored to address the challenge of ranking DMUs within DEA. To the best of our knowledge, no prior research has combined ensemble and DEA models, i.e., 'Ensemble-DEA', to generate, in this case, multiple BoD models through a randomization process across both the observations and the variables, thereby resolving the issue of ranking DMUs with high discriminatory power while preserving the full complexity of the dataset. This innovation maintains the integrity of the original data while ensuring that all relevant dimensions are considered. Although our methodology employs an ensemble structure, it does not involve the development or application of predictive models, nor does it incorporate a learning process in the conventional algorithmic sense. Rather, the innovation lies in extending the DEA framework through a double randomization procedure: constructing in our case multiple BoD models by

[4] Following the previous footnote, the values of optimal solution of model (1) are reciprocal of the solutions to the BoD model by Cherchye et al. (2004). Therefore, for model (1) we have values greater than or equal to one, while Cherchye's et al. (2004) model presents values smaller than or equal to one.

independently resampling both the set of DMUs and the set of indicators. This approach enhances the robustness and discriminatory power of the resulting composite indicators while preserving the interpretability and benchmarking features inherent to the BoD methodology.

3.1. Applying 'Ensemble-DEA' to calculate robust BoD composite indicators

The calculation of the BoD indicators through 'Ensemble-DEA' follows the next step-by-step detailed process:

- *Step 1. Data preparation and initial setup.* The first stage in our methodology involves thorough data preparation and initial setup. This phase is crucial as it lays the foundation for the entire analysis. The dataset is carefully examined for any inconsistencies or missing values, which are addressed to maintain the integrity and reliability of the analysis. Additionally, it is important to check for any variables that take negative values for some observations, as radially output-oriented DEA models like the BoD approach (1) is units invariant but not translation invariant. Another critical aspect of data preparation is identifying whether all indicators can be associated with 'good' indicators (where a higher value indicates better performance) or 'bad' indicators (where a higher value indicates worse performance). To ensure consistency, we transform all bad indicators into good indicators so that they point in the same direction of improvement. Initial parameters, such as the number of bootstrap samples p (usually 500 or 1,000) and the number of indicators to select at each iteration, are also defined. In particular, regarding the number of indicators to be randomly chosen for each subsample, we apply for simplicity the aforementioned heuristic rule proposed by Cooper et al. (2007), which has been widely used and cited in the literature. This rule is as follows: $n \geqslant \max\{m \cdot s, 3 \cdot (m+s)\}$. Given that the model (1) intrinsically assumes that all DMUs are using a single input equal to one for every unit, we have $m = 1$. Accordingly, we define the number of indicators to be used for each subsample, denoted as s^*, as the nearest integer to $w := \frac{n}{3} - 1$ that is greater than w. In this way, Cooper et al.'s rule of thumb is satisfied.
- *Step 2. Generation of bootstrap samples and application of random indicator selection.* Once the data is prepared, the next step is the generation of p bootstrap samples, denoted as $\aleph_q$, $q = 1, \ldots, p$. This involves creating multiple subsets of the original dataset by sampling with replacement following the ideas described in Section 2.2. Additionally, at each iteration, a random subset of indicators of size s^* is selected from the complete set of indicators. This approach ensures, at least with a high probability, that all dimensions of the dataset are considered if the value of parameter p is high, preventing any single indicator from dominating the analysis.
- *Step 3. Calculation of the BoD measures obtained for the bootstrap samples.* The third stage in our approach involves the construction of multiple models. For each bootstrap DMU sample and subset of indicators, we construct individual models using the BoD approach, model (1). Each BoD model evaluates the performance of the DMUs, providing a distinct perspective based on the specific observations and selected indicators for that particular sample. By applying the BoD model repeatedly across different bootstrap samples, we ensure that the analysis considers a wide range of potential data scenarios and interactions. Specifically, the model to be solved for unit $k \in \aleph_q$ would be as follows:

$$\begin{aligned} \phi_k^*(\aleph_q) = \quad & \max \quad && \phi_k \\ & s.t. \\ & && \sum_{i \in I_q} \lambda_{ki} y_{ir} \geqslant \phi_k y_{kr}, \quad r \in R_q \\ & && \sum_{i \in I_q} \lambda_{ki} = 1, \\ & && \lambda_{ki} \geqslant 0, \qquad i \in I_q, \end{aligned} \tag{2}$$

where R_q and I_q are the set of indices corresponding to the s^* indicators randomly selected (without replacement) from the original set of all indicators and the set of indices corresponding to the n randomly selected units (with replacement) from the original set of DMUs, respectively, for the subsample $\aleph_q$. Note that model (2) is an adaptation of model (1) where indicators (columns) and units (rows) correspond to elements that make up subsample $\aleph_q$.

- *Step 4. Aggregation of the BoD measures for each DMU.* The final stage involves aggregating the results from these multiple BoD models to form the final composite indicator of each DMU. For $k = 1, \ldots, n$, the composite indicator is defined as:

$$CI_k = \frac{1}{card\left(I_q^k\right)} \sum_{q=1}^{p} \sum_{k \in I_q} \phi_k^*(\aleph_q) \tag{3}$$

where $card\left(I_q^k\right)$ is the cardinal of the set $I_q^k := \{q / k \in I_q\}$. In other words, I_q^k is the set of indices such that the k-th unit is included in the q-th bootstrapped sample, regardless of the number of times that it is repeated in the subsamples where it appears. This means that $card\left(I_q^k\right)$ represents the number of bootstrap samples in which the k-th unit is included at least once. This calculation ensures that the composite indicator CI_k reflects the average performance of the k-th unit across all the bootstrap samples in which it appears.

3.2. Interpretation and properties of the aggregate composite indicator

The interpretation of CI_k is similar to that of a radial output-oriented efficiency score. Specifically, $100 \times (CI_k - 1)$ indicates the percentage by which all indicators should be increased, on average, across the p bootstrapped subsamples, for unit k to attain the BoD effective frontier. Therefore, the interpretation of the CI_k is the same as the output-oriented radial DEA model, but it reflects the mean performance across all the random scenarios considered in the analysis.

Additionally, the composite indicator CI_k has several important properties (see Pastor et al., 2022) inherited from the DEA radially output-oriented model, of which the BoD approach (models (1) and (2)) is a particular case. First, CI_k is bounded from below by one, where one represents that the DMU lays on the BoD effective frontier. This range indicates that the closer CI_k is to 1, the better the unit k is considered. Additionally, because the BoD models used in this approach are units' invariant, the composite indicator CI_k is also units' invariant. This means that the final scores are not affected by the scale of the indicators, making the comparison of units more consistent. Furthermore, the composite indicator CI_k satisfies the weak monotonicity property. In the case of BoD, this property ensures that if an observation improves or maintains its outputs, its performance score will not increase. Specifically, if one set of outputs $\widetilde{y}$ dominates another set y (i.e., $\widetilde{y} \geqslant y$), then $CI(\widetilde{y}) \leqslant CI(y)$. As a result, CI_k, which reflects the effectiveness of each unit, also upholds this property, meaning that any improvement in outputs will lead to an equal or better score, further enhancing the interpretability and robustness of the composite indicator in measuring performance. Also, CI_k meets the property of homogeneity of degree -1 in outputs. The homogeneity property makes easier the interpretation of

the results. If a DMU doubles all of its outputs, the performance improves in the same way by being halved. This provides a clear and intuitive understanding: a proportional improvement in outputs leads to a proportional reduction in the performance score (which means an improvement in performance, since the scores are greater than or equal to one, with one indicating the best performance).

Another key feature of the composite indicator is its capacity to capture variations in effectiveness across different scenarios, offering a more nuanced and comprehensive assessment of performance. By leveraging bootstrapped subsamples, the composite indicator accounts for the inherent variability in data, which enhances its robustness. This approach minimizes the risk of overfitting to a particular dataset and ensures that the results are generalizable across a wide range of contexts. Furthermore, the use of multiple subsamples introduces an element of stability, as the final score reflects the average performance over many different random combinations of indicators and units.

3.3. Benchmarking through peer identification

As commented in Section 2.1, the effectiveness of each DMU k is determined by comparing it against the benchmark frontier, which is constructed from the best performing units in the dataset. In the case of the BoD model, by solving the corresponding optimization model (2) we identify the set of peers for DMU k under evaluation, which are the units that form the convex combination upon which the evaluated DMU is projected through the radial measure. This benchmarking capability is particularly valuable, as it not only highlights ineffectiveness but also provides concrete targets and strategies for improvement by emulating the best practices of peer units.

We leverage on the benchmarking capabilities of DEA and extend them with our new approach, taking advantage of the evaluations performed with the many different randomly drawn subsamples. By doing so, we can replicate the peer identification and benchmarking capabilities of DEA, providing organizations with concrete strategies for improvement based on the practices of best performers.

The benchmarking process identifies peer units for each DMU by analyzing the intensity variables (lambda variables) from the BoD model solutions. For each DMU under evaluation, we identify relevant subsamples where the DMU appears. We then utilize the solutions to the model to identify the peers in each subsample. Peers are identified as DMUs with strictly positive lambda values. When handling multiple appearances, each peer DMU is counted only once. Peer information is aggregated across subsamples, and peer intensity percentages are calculated, representing the relevance of each peer. Higher intensity indicates a more crucial role as a benchmark for performance improvement.

In the subsequent Section, we demonstrate the effectiveness and applicability of the new approach in a real-world scenario.

4. Empirical illustration: Meeting the goals for sustainable development

4.1. Statistical sources, dataset and variables

The method developed in this paper is illustrated using data on Sustainable Development Goals (SDGs). SDGs are at the core of the 2030 Agenda for Sustainable Development established by United Nations in 2015, which is a plan of action for human well-being, planet protection from degradation and prosperity (United Nations, 2015). SDG data has been used previously in the context of BoD composite indicators (e.g., in Pereira et al., 2021). There are 17 dimensions (pillars) of SDGs that focus on such aims as, for example, no poverty and hunger, gender equality and quality of education. In turn, each pillar is composed of many indicators. The global database gives access to 248 indicators for countries, and areas or regions across the globe. We limit our analysis to the European Union (EU) using the set of indicators published by EUROSTAT (2023) in its 2023 edition. The goal is to focus on a set of countries sharing a common and binding institutional framework, e.g., the European climate law, and relatively comparable levels of economic development, hence having access to similar capabilities to achieve the SGDs. In the EU indicator set there are 102 variables for 17 SDGs, out of which 34 indicators are multipurpose; that is, they are used to track progress across multiple SDGs. We focus our analysis on year 2019, since it was the last year when data was available for most indicators.[5]

The empirical analysis required careful preparation of the data, including the removal of certain indicators due to missing observations and data adjustments. Firstly, many indicators had to be removed due to the absence of data for some countries; similarly, some indicators were dropped because they only contained data for specific year(s), excluding 2019. In addition, some indicators were excluded because they were not differentiated by country and only presented an EU average. One indicator was removed because it took negative values, which would be problematic for calculations using the BoD model, which is not translation invariant, see Section 3.2 on the properties of the aggregate performance indicators. The indicators in the dataset were of both 'good' output nature (where higher values indicate better performance) and 'bad' output nature (where higher values indicate worse performance). We transformed the bad indicators into good ones by calculating, for each country, the absolute value of the difference between the value of the bad output for the country in question less the maximum value of this bad output across all countries. Furthermore, some variables were provided in different units (for example, million euro and euro per inhabitant), and/or levels of disaggregation (for example, by sex or poverty status). For these variables we chose the most common unit of measurement and/or selected the aggregated values, respectively. Finally, some indicators were represented by multiple sub-indicators, so their average was considered in the analysis. As a result, the initial set of 102 indicators was reduced to 72, thereby completing a dataset for the countries and year considered, which were used in our computations. Table A.1 in the Appendix shows the final set of indicators, classified according to the type of variable ('good' or 'bad' output) and units of measurement.

4.2. Results

Following model (2), we apply the adaptation of ensemble techniques, i.e., 'Ensemble-DEA', with double randomization (DMUs and indicators) to the standard BoD model (1) considering the set of 72 SDGs indicators.[6] Table 1 ranks the 27 countries according to the average of their performance scores obtained for the total of 50,000 subsamples. We also used different subsamples sizes and found that similar results are obtained with 20,000 or 30,000 subsamples, which further enhance the robustness of our findings in terms of sensitivity in the variation in subsamples. We consider eight indicators in each subsample (see Step 1 in Section 3.1 and the adoption of Cooper et al.'s (2007) 'rule of thumb') and 27 countries. Furthermore, the subsampling process, which involves sampling with replacement, inherently leads to some countries appearing multiple times within the subsamples. This characteristic is not a flaw but rather a feature of the subsampling process, contributing

[5] Alternatively, we could focus our research on 2020, as it was another year with a significant amount of available data. However, since it was a pandemic year with potential distortions in the normal development of indicators, we decided not to consider this year in our analysis. Furthermore, from the initial sample of 28 EU countries in 2019, we were forced to remove the UK, which was in the process of exiting the EU and therefore there is a lack of data for many indicators in that specific year.

[6] The data and codes for estimation are hosted in the Zenodo repository (https://doi.org/10.5281/zenodo.15765362) and are linked to GitHub at https://github.com/mkapelko/Benefit-of-the-Doubt-Model-Enhancement-through-Ensemble-DEA.

Table 1
Comparison of BoD methods. Sustainable Development Goals, 2019.

Country	BoD-Ensemble-DEA		BoD- DEA*	BoD-SuperEffic.	BoD-Cross-Efficiency	
	Ranking	CI_k, eq. (3)			Benevolent	Aggressive
Sweden	1	1.00047	1	0.60159	1	0.61483
Finland	2	1.00193	1	0.72738	0.94387	0.51954
Denmark	3	1.00227	1	0.68283	1	0.55867
Luxembourg	4	1.00281	1	0.57834	0.99480	0.40451
Netherlands	5	1.00288	1	0.47973	1	0.58774
Ireland	6	1.00345	1	0.60271	1	0.40545
Slovenia	7	1.00539	1	0.81707	1	0.41449
Malta	8	1.00641	1	0.68329	1	0.35063
Germany	9	1.00855	1	0.33176	1	0.57985
Austria	10	1.00982	1	0.74828	0.99793	0.43518
France	11	1.01325	1	0.71519	0.99965	0.56991
Belgium	12	1.01383	1	0.77717	0.96580	0.45230
Cyprus	13	1.01399	1	0.45574	1	0.37995
Czechia	14	1.01437	1	0.77703	0.91937	0.40006
Estonia	15	1.01438	1	0.64018	1	0.41784
Lithuania	16	1.02033	1	0.75508	0.99945	0.30752
Slovakia	17	1.02061	1	0.82850	1	0.34156
Croatia	18	1.02631	1	0.78099	0.99895	0.34468
Latvia	19	1.02667	1	0.82918	0.98165	0.33945
Spain	20	1.02829	1	0.64847	1	0.41605
Hungary	21	1.03312	1	0.85118	1	0.30402
Italy	22	1.04321	1	0.46734	1	0.45562
Portugal	23	1.04703	1	0.74330	0.99884	0.37310
Bulgaria	24	1.06248	1	0.78661	0.99958	0.25922
Greece	25	1.06482	1	0.70225	0.99975	0.41725
Poland	26	1.07057	1	0.81718	0.93018	0.35079
Romania	27	1.10983	1	0.78276	0.94367	0.25850
Average		1.02471	1	0.68930	0.98791	0.41699
Stand. Dev.		0.02641	0	0.13303	0.02425	0.09927
Minimun		1.00047	1	0.33176	0.91937	0.25850
Maximum		1.10983	1	0.85118	1	0.61483

Source: Own elaboration.
Note: * Like the standard DEA model (1), 'BoD-Bootstrapping' and 'BoD-Order-m' also yield unitary efficiency scores. In the former case, this impedes the calculation of the distributions of the efficiency scores.

to the diversity of the subsamples and the overall robustness of the aggregation. As highlighted by Breiman (1996), when generating a bootstrap replicate of the same size as the original dataset, approximately 36.8 % of the cases can be expected to be repeated in the bootstrap sample. Finally, we calculate its mean score CI_k according to (3), and recover its benchmark peer(s) in each subsample: $\lambda^*_{ki} > 0$.

The results for the new approach (BoD-'Ensemble-DEA') show that the best and worst performing countries are Sweden and Romania with respective scores of 1.00047 and 1.10983. The average of all countries stands at 1.02471 with a standard deviation of 0.02641. Before discussing the relative performance of the countries with respect to the SDG indicators, we compare our results with those obtained using alternative methods reviewed in the introduction: 'BoD-DEA', corresponding to the standard DEA model (1), 'BoD-Superefficiency', 'BoD-Cross-Efficiency', implemented in its benevolent and aggressive approaches, 'BoD-Bootstrapping', and 'BoD-Order-m' (these methods measure effectiveness but for convenience we refer to them using the usual efficiency term). All these models are solved using the MATLAB toolbox developed by Álvarez et al. (2020). The only exception is the 'BoD-Order-m' method that is solved using the '*rcDEA*' package, Mergoni (2022).

4.2.1. Comparison with alternative methods

Compared to standard DEA, the ensemble methodology allows to overcome the lack of discriminatory power due to the curse of dimensionality. Following Cooper et al. (2007), with 27 countries and 72 indicators the usual rule determining the advised proportionality between observations and variables is not met, i.e., $n \geqslant \max\{m \cdot s, 3(m+s)\}$.[7] Under the BoD approach the number of inputs m can be considered as one (see footnote 3) and therefore, satisfying this condition would require using data on 219 countries. However, researchers are interested in assessing countries' performance across homogenous geopolitical environments like the EU, and therefore the condition cannot be met, finding that all countries are deemed best performers with a unitary score—see 'BoD-DEA' column.

The use of methods like 'BoD-Superefficiency' or 'BoD-Cross-Efficiency' does solve the ranking problem, but the attained scores are not interpretable in terms of performance improvements nor serve for benchmarking. Starting with the simple 'BoD-Superefficiency' model, the results indicate that Sweden, once removed from the sample when calculating the score, would have to radially reduce the value of the SDG indicators by 39.841 % ($=1 - 0.60159 \times 100$) to reach the reference hyperplane on the resulting frontier. Besides indicating the relative dominance of Sweden with large values in all indicators, not much can be learned from this result. For example, determining the peers on the new frontier is uninformative because this is a hypothetical exercise not intended to draw policy recommendations, i.e., why would stakeholders in Sweden be interested in these worse performing countries (including Cyprus)?[8] Similarly, the more complex 'BoD-Cross-Efficiency' methodology, either following the so-called benevolent or aggressive approach, offers an alternative ranking of countries. The method is based on the definition of the bilateral or 'peer-appraisal' (cross-efficiency) scores of

[7] Additionally, none of the other commonly referenced rules of thumb for balancing the number of DMUs with the number of variables are satisfied in this case.

[8] The reference peers for Sweden are Denmark ($\lambda_{ki} = 0.671$), France ($\lambda_{ki} = 0.099$), Cyprus ($\lambda_{ki} = 0.044$), Austria ($\lambda_{ki} = 0.114$), and Finland ($\lambda_{ki} = 0.717$).

an observation using the optimal weights of the remaining observations, which corresponds to the multipliers of the program dual to (1). Then, considering all n-1 bilateral cross-efficiencies, the benevolent and aggressive approaches are calculated by solving a program that either maximizes or minimizes the sum of all bilateral peer-appraisals, subject to the restrictions that: a) the self-appraisal (own efficiency) scores for each observation remains equal to the result obtained when solving (1), and b) no peer-appraisal score is greater than one. Balk et al. (2021) offer further details on the interpretation and calculation of cross-efficiency scores. However, as in the case of 'BoD-Superefficiency', from the cross-efficiency analysis one cannot derive concrete policy guidelines because it combines all bilateral cross-efficiencies, making it impossible to identify specific peers that can serve as benchmarks.

As for the 'BoD-Bootstrapping' method, it is not a valid methodological alternative for calculating scores different from one. Bootstrapping is based on the resampling of the original dataset by selecting a (large) number of independent subsamples where the observations are randomly replaced. However, as happens with the original sample where all observations have scores equal to one, in most bootstrapped samples all observations have again unitary score. While it is possible that an observation is randomly left out of a subsample, the average of all bootstrapped scores (500) effectively tends to one. This prevents the existence of any variability in the bootstrapped scores from which recover their distribution—notice that the mean of the score distribution is one and its variance zero, which impedes obtaining an estimate of the bias of the true value, and subsequently a bias-corrected estimator. The fact that the corrected estimators cannot be calculated explains why this method is not reported in Table 1. In the same vein, the 'BoD-Order-m' method runs into the same problems. This method models the benchmarking analysis as a stochastic process that compares the score obtained when solving (1) with the whole sample, against the expected efficiency scores that are obtained against subsamples of observations that obtain larger scores, i.e., those that dominate the observation under evaluation. In practice, the computations involve a simulation in which the performance of each observation is evaluated in many computational rounds (e.g., problem (1) is solved 500 times) using as reference a subsample of m randomly selected observations among those with equal or better performance levels on the s indicators. However, as with the bootstrapped method, even if an observation can be superefficient when randomly left out of a subsample, when the number of observations m and repetitions grow, the scores tend to one, and, once again, the method is uncapable of discriminating among observations—hence, we do not report this result in Table 1.[9]

Finally, notice that, even if this perfect realization of the curse of dimensionality were not observed, as in our empirical application, resampling techniques do not perform well in general when the proportion of best performers is large, resulting in a low variability of the resampled scores, either through Bootstrapping or Order-m methods. The reason is that these methods only resample across observations and not variables, and therefore the low proportionality of observations to the variables remains.

The above discussion summarizes why some of the most popular methods fail to solve the curse of dimensionality in a way that breaks the tie among efficient observations (e.g., 'BoD-Bootstrapping' and 'BoD-Order-m') or, if they do, the efficiency scores do not have a practical interpretation in terms of potential efficiency improvements, neither identify peers that can inform policy guidelines aimed at mirroring the best practices of these benchmarks (e.g., 'BoD-Superefficiency' and 'BoD-Cross-Efficiency').

In contrast, the ensemble method with double randomization proposed here provides a robust calculation of the performance scores solving the previous weaknesses. First, the average score can be clearly interpreted as the potential improvement in effectiveness that can be obtained if observations were to increase their output in the same proportion, what is explained in Section 4.2.2. Secondly, it is possible to identify and examine the most frequent peers on the frontier across the subsamples, as we comment in Section 4.2.3 and report in Table 2.

4.2.2. Performance

Considering the worst and best performing countries in Table 1, we see that Romania could increase its effectiveness by 10.983 % on average by expanding in the same proportion its SDG indicators, thereby reaching the BoD frontiers. Alternatively, Sweden is practically effective with an average efficiency score of $CI_{Sweden} = 1.00047$. The Kernel density distributions of the individual scores for these two countries are presented in Fig. 1.[10] The graph on the left panel (Fig. 1a) shows the absolute kernel density values of the performance scores (truncated at 1.2, for visual convenience), with the scale of the Romanian values portrayed in left vertical axis and that of the Swedish values in the right vertical axis. We see that, for Sweden, the density of its scores is very high around one (this country is effective in 98.1 % of the approximately 32,000 subsamples where it is present) while that of Romania is much lower (it is only effective 24.9 % of the times, thereby exhibiting a bimodal distribution), steadily decreasing after reaching a second maximum around its average score 1.11, and showing that this country is clearly ineffective when reaching SDG goals. In the case of Romania, the probability density for the performance scores accumulates almost entirely in the interval [1, 1.2], although this country presents scores greater than 5 in some simulations. In contrast, Sweden accumulates almost all its density in the interval [1, 1.01], with the worst result in all simulations being approximately 1.1. The difference between both distributions, once their values have been respectively normalized (divided) by their maximum value, is presented in the graph on the right panel (Fig. 1b). The same pattern can be visualized with both distributions converging in normalized value around 1.5. To confirm that the distributions are indeed statistically different we have performed three non-parametric tests (given that the efficiency scores are not normally distributed). First, the Mann-Whitney U test (or Wilcoxon rank-sum test), with a value of the statistic $U = 885{,}300{,}509.00$ (p-value = 0.0000) confirms that the distributions correspond to two independent groups. Similarly, the Kolmogorov-Smirnov (K-S) test, with a statistic $D = 0.7164$ (p-value = 0.0000), also returns that there are significant differences between them. Finally, applying Simar and Zelenyuk's (2006) test yields a statistic $L = 2{,}549.94$ (p-value = 0.0000), which once again confirms that they are statistically different from each other.[11]

As for the scores reported in Table 1, we see that like in any BoD analysis based on socioeconomic indicators, the ranking of countries is positively correlated with their value. Therefore, more socioeconomically developed and environmentally conscious countries stand at the top of the list, while the opposite is observed for less developed countries with weak welfare states and at the early stages of the green transition. The geographical divide between Northern and Southern EU countries on the one hand, and between Western and Eastern EU countries on the other hand, is clearly visible in Fig. 2. Scandinavian, Northern European and Benelux countries emerge as the best performers, while Eastern EU countries are the worst performers.

[9] Running alternative Order-m models (1) for different m sizes using the '*rcDEA*' software always yield unitary efficiency scores for all observations.

[10] The 'kdeplot' function (using the proposed parameters by default), from the 'seaborn' library, was used to represent the density functions. Seaborn is a Python data visualization library based on matplotlib (Waskom, 2021).

[11] These authors adapt Li's (1996) test to truncated data like BoD scores starting at one. The test does not assume any specific distribution for the efficiency scores and relies on bootstrap techniques to approximate the sampling distribution of the test statistic. To calculate the reported statistic we have performed 500 bootstrapped replications.

Table 2
Benchmark peers. BoD-'Ensemble-DEA'. Sustainable Development Goals, 2019, %.

Country	Belgium	Bulgaria	Czechia	Denmark	Germany	Estonia	Ireland	Greece	Spain	France	Croatia	Italy	Cyprus	Latvia	Lithuan.	Luxemb.	Hungary	Malta	Netherl.	Austria	Poland	Portugal	Romania	Slovenia	Slovakia	Finland	Sweden
Belgium	60.59	0.37	3.84	7.27	4.31	1.59	11.41	0.54	1.96	3.15	0.75	0.91	3.73	0.81	1.55	8.72	0.83	6.7	13.63	4.26	0.19	0.28	0.09	5.88	2.18	12.89	18.28
Bulgaria	2.32	29.18	2.28	4.65	1.56	13.62	9.73	0.79	2.61	1.72	6.26	0.8	5.68	7.56	3.93	9.72	1.28	19.73	4.93	4.69	0.23	1.92	1.51	8.38	1.65	9.45	20.35
Czechia	1.82	0.79	67.66	4.92	1.77	6.19	6.28	0.37	1.14	0.89	2.48	0.66	2.29	0.91	1.89	6.96	1.01	6.62	5.92	4.74	0.41	0.32	0.34	7.68	0.96	8.04	11.03
Denmark	0.97	0.08	1	89.66	0.74	0.93	2.26	0.05	0.21	0.53	0.28	0.21	0.77	0.32	0.51	3.21	0.13	1.66	2.51	1.17	0.04	0.08	0.03	2.65	0.74	3.4	5.92
Germany	1.23	0.05	1.03	4.95	83.55	0.34	2.77	0.2	0.63	1.64	0.22	0.72	0.55	0.2	0.23	2.29	0.14	1.43	4.93	2.4	0.17	0.14	0.11	0.97	0.32	3.17	7.46
Estonia	0.73	0.04	1.38	2.64	0.34	79.15	2.25	0.05	0.25	0.98	0.7	0.07	2.31	0.95	1.24	4.14	0.71	5.5	2.16	2.1	0.43	0.17	0.03	2.86	0.65	4.74	4.9
Ireland	0.73	0.04	1.44	2.16	0.53	1.25	87.17	0.06	0.22	0.36	0.63	0.09	1.48	0.47	0.53	4.06	0.31	3.81	2.19	1.04	0.1	0.12	0.03	2.45	0.79	3.58	5.13
Greece	3.8	1.8	3.78	8.73	4.76	5.47	13.8	32.66	4.81	2.61	3.37	4.14	11.26	2.45	1.82	10.11	1.11	21.05	12.38	4.37	0.33	2.84	1.16	9.94	4.19	6.91	21.44
Spain	2.18	0.28	1.03	4.9	9.01	1.36	10.4	1.08	63.98	6.05	1.41	5.56	3.1	0.75	0.94	3.42	0.5	5.69	7.43	3.08	0.31	0.28	0.53	2.27	0.89	5.21	12.49
France	2.39	0.31	3.42	4.7	9.45	0.79	6.81	0.45	3.1	72.6	0.69	2.21	1.92	0.17	0.8	3	0.35	3.48	8.32	3.02	0.13	0.35	0.22	2.78	1.02	5.13	11.94
Croatia	1.47	1.19	2.35	3.82	1.72	9.75	6.77	0.35	1.03	1.22	53.79	0.8	7.08	4.07	1.38	9.07	0.64	19.45	3.02	2.28	0.2	1.11	0.68	11.72	1.86	6.21	10.53
Italy	2.59	0.82	1.36	5.08	14.02	1.36	13.16	0.6	5.07	7.14	2.17	58.59	1.08	1.03	0.8	1.69	1.72	5.2	7.76	5.82	0.31	1.03	0.69	2.54	1.67	3.77	12.38
Cyprus	0.84	0.44	2.96	1.63	1	3.21	4.97	0.23	1.12	0.7	0.55	0.72	77.3	0.57	0.57	5.65	0.74	8.42	2.47	0.77	0.2	0.27	0.15	1.72	1.13	3.87	5.29
Latvia	0.87	0.84	1.32	2.77	0.45	14.01	5.48	0.22	0.71	1.6	3.66	0.4	7.16	54.31	4.23	8.55	1.31	13.83	1.47	2.48	0.62	0.59	0.48	6.13	0.81	6.07	11.7
Lithuania	1.04	1.4	1.44	3.91	1.14	9.48	6.18	0.2	1.1	0.89	4.22	0.35	6.87	5.33	56.34	10.29	1.11	13	1.97	2.92	0.2	0.81	0.56	8.1	0.5	6.39	11.17
Luxembourg	0.34	0.05	0.33	1.33	0.74	0.46	1.23	0.07	0.26	0.23	0.12	0.24	0.62	0.05	0.21	93.12	0.08	1.56	1.33	0.95	0.11	0.02	0.02	0.58	0.06	1.42	2.58
Hungary	4.55	1.82	7.92	10.95	3.01	9.8	12.44	0.77	2.76	2.64	7.94	1.96	5.86	3.42	5.9	11.26	35.94	19.49	7.68	5.77	1.69	0.8	0.57	13.23	4.8	9.42	15.82
Malta	0.52	0	1.14	0.74	0.42	0.4	1.93	0.11	0.16	0.27	0.02	0.25	0.73	0.03	0.04	0.82	0.13	91.27	2.41	0.25	0.15	0.01	0.04	0.54	0.4	1.36	3.11
Netherlands	0.76	0.09	0.66	2.53	2.73	0.13	2.68	0.12	0.57	0.87	0.17	0.3	0.5	0.15	0.22	2.19	0.02	1.01	89.78	0.79	0.04	0.12	0.06	1.31	0.27	2.97	5.2
Austria	2.56	0.06	2.64	10.27	3.38	2.65	6.69	0.43	0.59	1.32	0.76	0.63	2.45	0.61	1.09	7.48	0.56	5.77	7.49	70.72	0.35	0.27	0.06	5.82	0.96	7.94	11.31
Poland	6.14	2.63	9.09	7.66	11.73	8.19	19.26	1.72	6.27	8.21	7.6	4.71	4.91	1.9	4.49	5.21	5.43	10.69	13.55	5.94	27.08	1.31	1.07	13.96	6.93	10.85	21.13
Portugal	4.07	1.02	2.98	16.81	3.53	8.39	10.66	1	4.68	3.8	5.68	2.91	6.08	6.55	8.34	14.46	1.91	22.11	12.7	5.18	0.71	26.59	0.99	12.82	1.85	17.02	25.75
Romania	2.34	3.43	4.4	7.53	3.06	12.72	12.9	1.77	5.33	3.55	4.33	3.25	6.17	6.33	3.48	7.85	5.44	17.05	5.21	4.91	2.23	3.17	24.88	4.68	3.56	7.81	24.42
Slovenia	0.93	0.38	1.04	1.85	0.83	2.85	3.2	0.14	0.48	0.96	0.87	0.25	2.15	1.1	1.01	3.3	0.21	4.69	2.08	1.05	0.17	0.23	0.18	83.45	0.42	4.04	6.45
Slovakia	1.53	2.26	3.27	3.83	1.21	9.78	7.55	0.47	0.93	0.84	5.83	0.54	7.26	5.6	6.61	10.34	0.87	14.17	3.33	3.28	0.54	0.82	0.37	14.53	50.75	9.41	10.72
Finland	0.36	0.09	0.85	2.53	0.26	0.8	1.17	0.08	0.07	0.28	0.41	0.07	0.44	0.28	0.3	1.72	0.13	1.15	1.21	0.67	0.12	0.05	0.04	1.35	0.29	92.74	4.21
Sweden	0.21	0.01	0.26	0.61	0.27	0.12	0.57	0.04	0.07	0.16	0.06	0.04	0.24	0.03	0.06	0.73	0.07	0.31	0.49	0.24	0.04	0.01	0	0.25	0.06	0.47	98.13
Average	1.82	0.78	2.43	4.95	3.15	4.83	7.02	0.46	1.77	2.02	2.35	1.26	3.57	1.99	2.01	6.01	1.03	8.98	5.33	2.85	0.39	0.66	0.39	5.58	1.50	6.21	11.57
Stand. Dev.	1.47	0.93	2.14	3.70	3.75	4.71	4.85	0.48	1.92	2.16	2.53	1.57	2.95	2.38	2.26	3.77	1.40	7.18	3.80	1.89	0.50	0.84	0.42	4.66	1.65	3.78	6.77
Maximum	6.14	3.43	9.09	16.81	14.02	14.01	19.26	1.77	6.27	8.21	7.94	5.56	11.26	7.56	8.34	14.46	5.44	22.11	13.55	5.94	2.23	3.17	1.51	14.53	6.93	17.02	25.75
Minimum	0.21	0.00	0.26	0.61	0.26	0.12	0.57	0.04	0.07	0.16	0.02	0.04	0.24	0.03	0.04	0.73	0.02	0.31	0.49	0.24	0.04	0.01	0.00	0.25	0.06	0.47	2.58

Note: For each evaluated country (rows), the columns show the number of times—in percentage terms over the number of random samples where the DMU is present—that a country is a benchmark peer. The main diagonal reflects the number of times that a country has a score equals one being benchmark to itself. More intense colour shades indicate bigger percentages (green) and lower percentages (red).

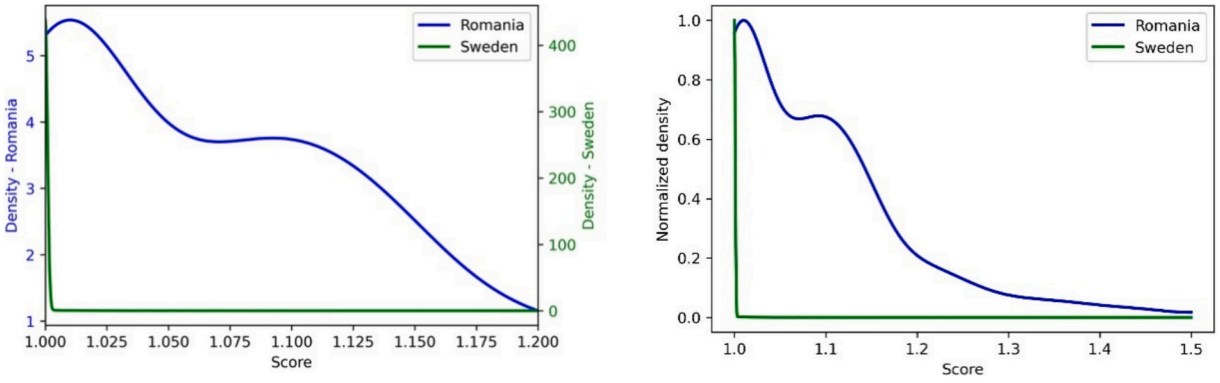

Fig. 1. Kernel density functions of scores using the Bod-'Ensemble-DEA' method. Fig. 1a (left), absolute density values. Fig. 2a (left) Right, normalized density values.

4.2.3. *Benchmark peers*

Relevant for benchmarking is the identification of the peers for each country on the BoD frontiers. As discussed in the methodological Section 2.1, we are interested in identifying the countries that are benchmarks when evaluating the effectiveness of the country under evaluation; i.e., those with $\lambda^*_{ki} > 0$ when solving program (2). Reading Table 2 by columns, we report the number of times (in percentage) that a country serves as benchmark for each evaluated row country, considering the 50,000 subsamples.

Taking the first entry as example, Belgium identifies itself as benchmark peer in 60.59 % of the total number of runs, implying that it is one of the best performers in those subsamples with $\phi^*_k(\aleph_q) = 1$ as optimal solution to program (2). This shows that the main diagonal identifies the number of times that the country under evaluation is effective by defining the BoD frontier. Logically, Sweden, the best country, is its own peer most of the times, 98.1 %. On the contrary, still within the main diagonal, Romania is the country that is deemed underperforming most frequently, as it only comes out efficient in 24.9 % of the subsamples—followed by Poland with 27.1 %.

Coming back to Belgium in the first column we observe that despite being a relatively effective country—it ranks 12th in CI_k, it is not generally identified as peer. The last four rows present descriptive statistics of peer frequency by column excluding the main diagonal, i.e., leaving out self-appraisals in the spirit of cross-efficiency methods. On average, Belgium is only benchmark peer in 1.82% of the samples, which stands among the lowest observed values. On the contrary, the best country, Sweden, serves as benchmark in 11.6% of the subsamples, followed by Malta (9.0%, ranking 8th in CI_k) and Ireland (7.0%, ranking 6th in CI_k). In this way, the BoD methodology allows identifying, for each evaluated country, the peers that serve as reference for performance improvements. Excluding self-appraisals in the main diagonal, Sweden and Scandinavian countries emerge as role models, but countries that have experienced fast economic progress in the last decade, like Ireland, or small countries in terms of size and population but offering specialized financial and fiscal services like Malta or Luxembourg are also frequently identified as peers. For underperforming European countries, choosing a role model should consider the likelihood that they can mirror the best practices of their peers in terms of the SDGs

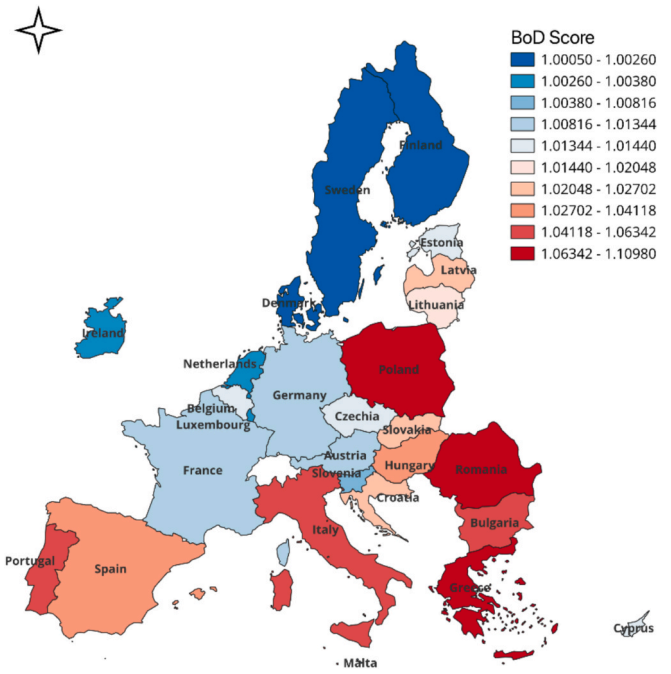

Fig. 2. BoD-'Ensemble-DEA' scores, CI_k—eq. (3). Sustainable Development Goals, 2019. .
Source: Own elaboration

indicators, as it is not always feasible to adopt the socioeconomic and institutional background of some of the leading countries.

Finally, it is possible to resort to Table 2 to determine which country is more relevant in a bilateral comparison taking their respective frequency as reference peers. For example, considering the best and worst performing countries, Sweden and Romania, the former is peer to the latter in 24.4% of the samples. On the contrary, Romania is never a peer to Sweden. A way to establish a ranking of countries in bilateral comparisons is to solve the linear ordering problem of the peer matrix (Ceberio et al., 2015). Given the interpretative layout of the peer matrix with benchmark countries in columns and evaluated countries in rows, the method consists of finding a simultaneous permutation of the rows and columns of Table 2, such that the sum of the entries below the main diagonal is maximized (or equivalently, the sum of the entries above the main diagonal is minimized). This reorders the countries according to their ranking when serving as peers for the remaining $n - 1$ countries. The resulting matrix is presented in Table A.2 of the Appendix. Sweden is the most frequent peer in any bilateral comparison, followed by Luxemburg, Finland and The Netherlands. In descending order, Greece, Poland and Romania close the ranking in the last positions. This method provides a robust ranking of peers based on a novel ensemble approach that combines repeated BoD estimations across random subsets of indicators and DMUs. By aggregating these results, it delivers reliable performance assessments and benchmarking insights that can guide policy strategies aimed at improving the current situation of countries in terms of SDG performance.

5. Conclusions: implications, limitations, and future research

5.1. Summary and implications

In this study we introduce 'Ensemble-DEA', a methodology integrating DEA and ensemble methods that is used to construct composite indicators within the BoD approach, specifically utilizing double randomization in both observations and variables (indicators). Our approach addresses critical issues associated with traditional methods, particularly the curse of dimensionality, which hampers the discriminatory power and interpretability of performance scores when the number of indicators is high relative to the number of observations.

Our methodology presents significant advantages over existing techniques. First, it preserves the integrity of the dataset by maintaining all available information, thus ensuring no loss of critical data. Second, by employing our approach, we enhance the robustness and reliability of the composite indicator. This allows for a more credible evaluation of DMUs, providing scores with high discriminatory power and enabling certain and accurate rankings. Third, our method retains the interpretability of traditional DEA scores while allowing the identification of benchmark peers, making the results clearer for policymakers.

The application of our method to a real-world dataset from EUROSTAT, encompassing 27 countries and 72 variables related to the Sustainable Development Goals, has illustrated its relevance and usefulness, highlighting theoretical (methodological) and practical implications. The methodological implications underscore the robustness and superiority of our method in handling complex datasets with numerous indicators as compared to standard approaches developed in the prior literature. This implies that the new approach is currently the only available solution for obtaining reliable composite indicators from complex datasets that have a clear efficiency or effectiveness interpretation and provide valuable benchmarking information, while preserving the integrity of the dataset. On the one hand, popular methods like Superefficiency or Cross-Efficiency allow overcoming the curse of dimensionality and ranking problem, but the efficiency scores and peer analysis do not convey sensible information in terms of managerial and policymaking. In particular, they are not interpretable in terms of performance improvements and do not allow to identify benchmarks. On the other hand, more sophisticated methods based on resampling like Bootstrapping or Order-m suffer from the same disadvantages that standard DEA methods by yielding unitary scores for all observations. The reason is that they only randomize by observations but not variables. Hence, it is impossible to discriminate between observations and indicate which units are the best and which are the worst performers. Contrarily to the above, the new method offers reliable information about: 1) the relative performance of countries through the proposed composite indicator that summarizes the scores obtained from numerous subsamples, and 2) country-specific peers that serve as benchmarks to improve the performance when meeting the goals for sustainable development.

Regarding the practical implications, from the analysis we observe that performance in meeting SDGs within the EU is highly correlated with the level of socioeconomic development and environmental consciousness, with Scandinavian, Northern European and Benelux countries performing best, while Eastern European countries lag in the efficiency and peer rankings. Even within the BoD approach that offers the most favourable comparison, this indicates that to improve their performance levels the latter countries must adopt policies aimed at reducing poverty levels, provide good health and wellbeing, quality education, gender equality, etc. Clearly the largest the performance score, the more efforts should be placed in simultaneously improving all SDGs. This is particularly relevant for Romania and Bulgaria, which have the lowest scores in SDGs realization. This also implies that Eastern European countries should consider undertaking benchmarking exercises to learn how Scandinavian, Northern European and Benelux societies implement sustainability principles related to the SDGs. In this context, sharing best practices and fostering collaboration between countries could facilitate this process.

5.2. Limitations and future research

While our novel methodology has shown promising results, several avenues for future research remain open. First, in this study, we relied solely on the rule of thumb proposed by Cooper et al. (2007) for determining the number of indicators in each subsample. However, there are other heuristics available in the literature that could be

explored to assess their impact on improving the discriminatory power of our model and robustness of findings. Future research could compare our results with those obtained under alternative rules and optimization strategies for variable selection within the DEA framework. Second, in our current approach, the identification of benchmark peers is based on whether the intensity variables are strictly greater than zero. A promising direction for future research could involve using the exact values of these intensity variables to provide more granular insights into the relevance and contribution of each peer. This refinement could lead to a more detailed (e.g., weighted) benchmarking process and offer more precise guidance for performance improvements. Third, extending the method beyond the BoD approach to a more general and conventional DEA production context including both inputs and outputs is a straightforward extension of this study. Fourth, another methodological refinement could consist in extending the robustness and sensitivity analysis of our method towards perturbation and variations in the data variables. For example, the indicators could be altered with some margins, in line with the literature on robust DEA (e.g., Arabmaldar et al., 2023). Fifth, from the empirical point of view, the analysis of effectiveness of SDGs implementation over time could provide insights into performance changes and potential catching up between countries. Finally, further empirical validation of our method across different sectors and contexts is necessary to establish its generalizability. Applying the methodology to diverse datasets, such as healthcare, education, and finance, would help in understanding its broader applicability and potential limitations.

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.

Acknowledgements

The financial support for this article from the National Science Centre in Poland (grant no. 2023/49/B/HS4/02991) is gratefully acknowledged. Some calculations were made at the Wroclaw Centre for Networking and Supercomputing (www.wcss.wroc.pl), computing grant no. 286.

Appendix A. Additional tables

.

Table A1
SDGs indicators used in the empirical analysis.

Goals	Indicators	Type of output	Units of measurement
Goal 1. No poverty	SDG_01_10. People at risk of poverty or social exclusion	Bad	Thousand persons
	SDG_01_20. Persons at risk of monetary poverty after social transfers	Bad	Thousand persons
	SDG_01_31. Severe material and social deprivation rate	Bad	Thousand persons
	SDG_01_40. People living in households with very low work intensity	Bad	Thousand persons
	SDG_01_41. In work at-risk-of-poverty rate	Bad	Percentage
	SDG_01_50. Housing cost overburden rate	Bad	Percentage
Goal 2. Zero hunger	SDG_02_20. Agricultural factor income per annual work unit	Good	Euro per annual work unit
	SDG_02_30. Government support to agricultural research and development	Good	Million euro
	SDG_02_40. Area under organic farming	Good	Percentage of total utilised agricultural area
	SDG_02_60. Ammonia emissions from agriculture	Bad	Tonne
Goal 3. Good health and wellbeing	SDG_03_11. Healthy life years at birth	Good	Years
	SDG_03_20. Share of people with good or very good perceived health	Good	Percentage
	SDG_03_41. Standardised death rate due to tuberculosis, HIV and hepatitis	Bad	Rate
	SDG_03_42. Standardised preventable and treatable mortality	Bad	Rate
	SDG_03_60. Self-reported unmet need for medical examination and care	Bad	Percentage
Goal 4. Quality education	SDG_04_10. Early leavers from education and training	Bad	Percentage
	SDG_04_20. Tertiary educational attainment	Good	Percentage
	SDG_04_31. Participation in early childhood education (children aged 3 and over)	Good	Percentage
	SDG_04_60. Adult participation in learning in the past four weeks	Good	Percentage
Goal 5. Gender equality	SDG_05_30. Gender employment gap	Bad	Percentage point
	SDG_05_40. Persons outside the labour force due to caring responsibilities	Bad	Percentage of total population
	SDG_05_50. Seats held by women in national parliaments and governments	Good	Percentage
	SDG_05_60. Positions held by women in senior management positions	Good	Percentage
Goal 6. Clean water and sanitation	SDG_06_10. Population having neither a bath, nor a shower, nor indoor flushing toilet in their household	Bad	Percentage
	SDG_06_60. Water exploitation index	Bad	Percentage
Goal 7. Affordable and clean energy	SDG_07_10. Primary energy consumption	Bad	Million tonnes of oil equivalent
	SDG_07_11. Final energy consumption	Bad	Million tonnes of oil equivalent

(continued on next page)

Table A1 (*continued*)

Goals	Indicators	Type of output	Units of measurement
	SDG_07_20. Final energy consumption in households per capita	Bad	Kilogram of oil equivalent
	SDG_07_30. Energy productivity	Good	Purchasing power standard per kilogram of oil equivalent
	SDG_07_40. Share of renewable energy in gross final energy consumption	Good	Percentage
	SDG_07_50. Energy import dependency	Bad	Percentage
	SDG_07_60. Population unable to keep home adequately warm	Bad	Percentage
Goal 8. Decent work and economic growth	SDG_08_10. Real GDP per capita	Good	Euro per capita
	SDG_08_20. Young people neither in employment nor in education and training	Bad	Percentage of total population
	SDG_08_30. Employment rate	Good	Percentage of total population
	SDG_08_40. Long-term unemployment rate	Bad	Rate
	SDG_08_60. Fatal accidents at work per 100 000 workers	Bad	Rate
Goal 9. Industry, innovation and infrastructure	SDG_09_10. Gross domestic expenditure on R&D	Good	Percentage of gross domestic product
	SDG_09_30. R&D personnel	Good	Percentage of population in the labour force
	SDG_09_40. Patent applications to the European Patent Office	Good	Number
	SDG_09_50. Share of buses and trains in inland passenger transport	Good	Percentage
	SDG_09_70. Air emission intensity from industry	Bad	Grams per euro
Goal 10. Reduced inequalities	SDG_10_10. Purchasing power adjusted GDP per capita	Good	Real expenditure per capita
	SDG_10_30. Relative median at-risk-of-poverty gap	Bad	Percentage
	SDG_10_41. Income distribution	Bad	Ratio
	SDG_10_50. Income share of the bottom 40 % of the population	Good	Percentage
	SDG_10_60. Asylum applications	Good	Number
Goal 11. Sustainable cities and communities	SDG_11_11. Severe housing deprivation rate	Bad	Percentage
	SDG_11_20. Population living in households considering that they suffer from noise	Bad	Percentage
	SDG_11_40. Road traffic deaths	Bad	Number
	SDG_11_52. Premature deaths due to exposure to fine particulate matter	Bad	Number
	SDG_11_60. Recycling rate of municipal waste	Good	Percentage
Goal 12. Responsible consumption and production	SDG_12_21. Raw material consumption	Bad	Thousand tonnes
	SDG_12_30. Average CO2 emissions per km from new passenger cars	Bad	Emissions per km
	SDG_12_41. Circular material use rate	Good	Percentage
	SDG_12_61. Gross value added in environmental goods and services sector	Good	Million euro
Goal 13. Climate action	SDG_13_10. Net greenhouse gas emissions	Bad	Tonnes per capita
	SDG_13_40. Climate related economic losses	Bad	Million euro
	SDG_13_60. Population covered by the Covenant of Mayors for Climate & Energy signatories	Good	Million persons
Goal 14. Life below water	SDG_14_40. Bathing sites with excellent water quality	Good	Number
Goal 15. Life on land	No indicators with available data		
Goal 16. Peace, justice and strong institutions	SDG_16_10. Standardised death rate due to homicide	Bad	Rate
	SDG_16_20. Population reporting occurrence of crime, violence or vandalism in their area	Bad	Percentage
	SDG_16_30. General government total expenditure on law courts	Good	Million euro
	SDG_16_40. Perceived independence of the justice system	Good	Percentage
	SDG_16_50. Corruption Perceptions Index	Good	Number
	SDG_16_60. Population with confidence in EU institutions	Good	Percentage
Goal 17. Partnerships for the goals	SDG_17_10. Official development assistance as share of gross national income	Good	Percentage
	SDG_17_20. EU financing to developing countries	Good	Value in current prices
	SDG_17_30. EU imports from developing countries	Good	Million euro
	SDG_17_40. General government gross debt	Bad	Million euro
	SDG_17_50. Share of environmental taxes in total tax revenues	Bad	Share
	SDG_17_60. High-speed internet coverage	Good	Percentage of households

Table A2
Ranking of benchmarks solving the linear ordering problem of the peer matrix. BoD-'Ensemble-DEA'. Sustainable Development Goals, 2019, %.

Rank	Country	1 Sweden	2 Luxemb.	3 Finland	4 Netherl.	5 Malta	6 Ireland	7 Slovenia	8 Denmark	9 Estonia	10 Germany	11 Cyprus	12 Austria	13 Czechia	14 France	15 Latvia	16 Croatia	17 Lithuan.	18 Slovakia	19 Belgium	20 Italy	21 Spain	22 Hungary	23 Portugal	24 Bulgaria	25 Greece	26 Poland	27 Romania
1	Sweden	98.13	0.73	0.47	0.49	0.31	0.57	0.25	0.61	0.12	0.27	0.24	0.24	0.26	0.16	0.03	0.06	0.06	0.06	0.21	0.04	0.07	0.07	0.01	0.01	0.04	0.04	0.00
2	Luxemb.	2.58	93.12	1.42	1.33	1.56	1.23	0.58	1.33	0.46	0.74	0.62	0.95	0.33	0.23	0.05	0.12	0.21	0.06	0.34	0.24	0.26	0.08	0.02	0.05	0.07	0.11	0.02
3	Finland	4.21	1.72	92.74	1.21	1.15	1.17	1.35	2.53	0.80	0.26	0.44	0.67	0.85	0.28	0.28	0.41	0.30	0.29	0.36	0.07	0.07	0.13	0.05	0.09	0.08	0.12	0.04
4	Netherland	5.20	2.19	2.97	89.78	1.01	2.68	1.31	2.53	0.13	2.73	0.50	0.79	0.66	0.87	0.15	0.17	0.22	0.27	0.76	0.30	0.57	0.02	0.12	0.09	0.12	0.04	0.06
5	Malta	3.11	0.82	1.36	2.41	91.27	1.93	0.54	0.74	0.40	0.42	0.73	0.25	1.14	0.27	0.03	0.02	0.04	0.40	0.52	0.25	0.16	0.13	0.01	0.00	0.11	0.15	0.04
6	Ireland	5.13	4.06	3.58	2.19	3.81	87.17	2.45	2.16	1.25	0.53	1.48	1.04	1.44	0.36	0.47	0.63	0.53	0.79	0.73	0.09	0.22	0.31	0.12	0.04	0.06	0.10	0.03
7	Slovenia	6.45	3.30	4.04	2.08	4.69	3.20	83.45	1.85	2.85	0.83	2.15	1.05	1.04	0.96	1.10	0.87	1.01	0.42	0.93	0.25	0.48	0.21	0.23	0.38	0.14	0.17	0.18
8	Denmark	5.92	3.21	3.40	2.51	1.66	2.26	2.65	89.66	0.93	0.74	0.77	1.17	1.00	0.53	0.32	0.28	0.51	0.74	0.97	0.21	0.21	0.13	0.08	0.08	0.05	0.04	0.03
9	Estonia	4.90	4.14	4.74	2.16	5.50	2.25	2.86	2.64	79.15	0.34	2.31	2.10	1.38	0.98	0.95	0.70	1.24	0.65	0.73	0.07	0.25	0.71	0.17	0.04	0.05	0.43	0.03
10	Germany	7.46	2.29	3.17	4.93	1.43	2.77	0.97	4.95	0.34	83.55	0.55	2.40	1.03	1.64	0.20	0.22	0.23	0.32	1.23	0.72	0.63	0.14	0.14	0.05	0.20	0.17	0.11
11	Cyprus	5.29	5.65	3.87	2.47	8.42	4.97	1.72	1.63	3.21	1.00	77.30	0.77	2.96	0.70	0.57	0.55	0.57	1.13	0.84	0.72	1.12	0.74	0.27	0.44	0.23	0.20	0.15
12	Austria	11.31	7.48	7.94	7.49	5.77	6.69	5.82	10.27	2.65	3.38	2.45	70.72	2.64	1.32	0.61	0.76	1.09	0.96	2.56	0.63	0.59	0.56	0.27	0.06	0.43	0.35	0.06
13	Czechia	11.03	6.96	8.04	5.92	6.62	6.28	7.68	4.92	6.19	1.77	2.29	4.74	67.66	0.89	0.91	2.48	1.89	0.96	1.82	0.66	1.14	1.01	0.32	0.79	0.37	0.41	0.34
14	France	11.94	3.00	5.13	8.32	3.48	6.81	2.78	4.70	0.79	9.45	1.92	3.02	3.42	72.60	0.17	0.69	0.80	1.02	2.39	2.21	3.10	0.35	0.35	0.31	0.45	0.13	0.22
15	Latvia	11.70	8.55	6.07	1.47	13.83	5.48	6.13	2.77	14.01	0.45	7.16	2.48	1.32	1.60	54.31	3.66	4.23	0.81	0.87	0.40	0.71	1.31	0.59	0.84	0.22	0.62	0.48
16	Croatia	10.53	9.07	6.21	3.02	19.45	6.77	11.72	3.82	9.75	1.72	7.08	2.28	2.35	1.22	4.07	53.79	1.38	1.86	1.47	0.80	1.03	0.64	1.11	1.19	0.35	0.20	0.68
17	Lithuania	11.17	10.29	6.39	1.97	13.00	6.18	8.10	3.91	9.48	1.14	6.87	2.92	1.44	0.89	5.33	4.22	56.34	0.50	1.04	0.35	1.10	1.11	0.81	1.40	0.20	0.20	0.56
18	Slovakia	10.72	10.34	9.41	3.33	14.17	7.55	14.53	3.83	9.78	1.21	7.26	3.28	3.27	0.84	5.60	5.83	6.61	50.75	1.53	0.54	0.93	0.87	0.82	2.26	0.47	0.54	0.37
19	Belgium	18.28	8.72	12.89	13.63	6.70	11.41	5.88	7.27	1.59	4.31	3.73	4.26	3.84	3.15	0.81	0.75	1.55	2.18	60.59	0.91	1.96	0.83	0.28	0.37	0.54	0.19	0.09
20	Italy	12.38	1.69	3.77	7.76	5.20	13.16	2.54	5.08	1.36	14.02	1.08	5.82	1.36	7.14	1.03	2.17	0.80	1.67	2.59	58.59	5.07	1.72	1.03	0.82	0.60	0.31	0.69
21	Spain	12.49	3.42	5.21	7.43	5.69	10.40	2.27	4.90	1.36	9.01	3.10	3.08	1.03	6.05	0.75	1.41	0.94	0.89	2.18	5.56	63.98	0.50	0.28	0.28	1.08	0.31	0.53
22	Hungary	15.82	11.26	9.42	7.68	19.49	12.44	13.23	10.95	9.80	3.01	5.86	5.77	7.92	2.64	3.42	7.94	5.90	4.80	4.55	1.96	2.76	35.94	0.80	1.82	0.77	1.69	0.57
23	Portugal	25.75	14.46	17.02	12.70	22.11	10.66	12.82	16.81	8.39	3.53	6.08	5.18	2.98	3.80	6.55	5.68	8.34	1.85	4.07	2.91	4.68	1.91	26.59	1.02	1.00	0.71	0.99
24	Bulgaria	20.35	9.72	9.45	4.93	19.73	9.73	8.38	4.65	13.62	1.56	5.68	4.69	2.28	1.72	7.56	6.26	3.93	1.65	2.32	0.80	2.61	1.28	1.92	29.18	0.79	0.23	1.51
25	Greece	21.44	10.11	6.91	12.38	21.05	13.80	9.94	8.73	5.47	4.76	11.26	4.37	3.78	2.61	2.45	3.37	1.82	4.19	3.80	4.14	4.81	1.11	2.84	1.80	32.66	0.33	1.16
26	Poland	21.13	5.21	10.85	13.55	10.69	19.26	13.96	7.66	8.19	11.73	4.91	5.94	9.09	8.21	1.90	7.60	4.49	6.93	6.14	4.71	6.27	5.43	1.31	2.63	1.72	27.08	1.07
27	Romania	24.42	7.85	7.81	5.21	17.05	12.90	4.68	7.53	12.72	3.06	6.17	4.91	4.40	3.55	6.33	4.33	3.48	3.56	2.34	3.25	5.33	5.44	3.17	3.43	1.77	2.23	24.88
	Average	11.57	6.01	6.21	5.33	8.98	7.02	5.58	4.95	4.83	3.15	3.57	2.85	2.43	2.02	1.99	2.35	2.01	1.50	1.82	1.26	1.77	1.03	0.66	0.78	0.46	0.39	0.39
	Stand. Dev.	6.77	3.77	3.78	4.09	7.18	4.85	4.66	3.70	4.71	3.75	2.95	1.89	2.14	2.16	2.38	2.53	2.26	1.65	1.46	1.57	1.92	1.40	0.84	0.93	0.48	0.50	0.42
	Maximum	25.75	14.46	17.02	13.63	22.11	19.26	14.53	16.81	14.01	14.02	11.26	5.94	9.09	8.21	7.56	7.94	8.34	6.93	6.14	5.56	6.27	5.44	3.17	3.43	1.77	2.23	1.51
	Minimum	2.58	0.73	0.47	0.49	0.31	0.57	0.25	0.61	0.12	0.26	0.24	0.24	0.26	0.16	0.03	0.02	0.04	0.06	0.52	0.04	0.07	0.02	0.01	0.00	0.04	0.04	0.00

Note: For each evaluated country (rows), the columns show the number of times— in percentage terms over the number of random samples where the DMU is present —that a country is benchmark peer. More intense colour shades indicate bigger percentages (green) and lower percentages (red) (the main diagonal is excluded since it does not contain bilateral information but self-appraisals).

Data availability

The links to the repositories containing the programming codes and dataset have been included in the manuscript.

References

Adler, N., & Golany, B. (2001). Evaluation of deregulated airline networks using data envelopment analysis combined with principal component analysis with an application to Western Europe. *European Journal of Operational Research, 132*(2), 260–273. https://doi.org/10.1016/S0377-2217(00)00150-8

Aldamak, A., & Zolfaghari, S. (2017). Review of efficiency ranking methods in data envelopment analysis. *Measurement, 106*, 161–172. https://doi.org/10.1016/j.measurement.2017.04.028

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. https://doi.org/10.1023/A:1018968909638

Álvarez, I. C., Barbero, J., & Zofío, J. L. (2020). A Data Envelopment Analysis Toolbox for MATLAB. *Journal of Statistical Software, 95*(3), 1–49. https://doi.org/10.18637/jss.v095.i03

Andersen, P., & Petersen, N. C. (1993). A procedure for ranking efficient units in data envelopment analysis. *Management Science, 39*(10), 1261–1264. https://doi.org/10.1287/mnsc.39.10.1261

Angiz, M. Z., Mustafa, A., & Kamali, M. J. (2013). Cross-ranking of decision making units in data envelopment analysis. *Applied Mathematical Modelling, 37*(1–2), 398–405. https://doi.org/10.1016/j.apm.2012.02.038

Aparicio, J., & Kapelko, M. (2019). Enhancing the measurement of composite indicators of corporate social performance. *Social Indicators Research, 114*(2), 807–826. https://doi.org/10.1007/s11205-018-02052-1

Aparicio, J., Kapelko, M., & Monge, J. F. (2020). A well-defined composite indicator: An application to corporate social responsibility. *Journal of Optimization Theory and Applications, 186*(1), 299–323. https://doi.org/10.1007/s10957-020-01701-1

Arabmaldar, A., Sahoo, B. K., & Ghiyasi, M. (2023). A generalized robust data envelopment analysis model based on directional distance function. *European Journal of Operational Research, 311*(2), 617–632. https://doi.org/10.1016/j.ejor.2023.05.005

Balk, B. M., De Koster, R., Kaps, C., & Zofío, J. L. (2021). An evaluation of cross-efficiency methods: With an application to warehouse performance. *Applied Mathematics and Computation, 406*, Article 126261. https://doi.org/10.1016/j.amc.2021.126261

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. https://doi.org/10.1287/mnsc.30.9.1078

Banker, R. D., Charnes, A., Cooper, W. W., Swarts, J., & Thomas, D. (1989). An introduction to data envelopment analysis with some of its models and their uses. *Research in Governmental and Nonprofit Accounting, 5*(1), 125–163.

Barbero, J., Zabala-Iturriagagoitia, J. M., & Zofío, J. L. (2021). Is more always better? on the relevance of decreasing returns to scale on innovation. *Technovation, 107*, Article 102314. https://doi.org/10.1016/j.technovation.2021.102314

Bardhan, I., Bowlin, W. F., Cooper, W. W., & Sueyoshi, T. (1996). Models and measures for efficiency dominance in DEA - Part I: Additive models and MED measures. *Journal of the Operations Research Society of Japan, 39*(3), 322–332. https://doi.org/10.15807/jorsj.39.322

Benítez-Peña, S., Bogetoft, P., & Morales, D. R. (2020). Feature selection in Data Envelopment Analysis: A mathematical optimization approach. *Omega, 96*, Article 102068. https://doi.org/10.1016/j.omega.2019.05.004

Breiman, L. (1996). Bagging predictors. *Machine learning, 24*, 123–140.

Ceberio, J., Mendiburu, A., & Lozano, J. A. (2015). The linear ordering problem revisited. *European Journal of Operational Research, 241*(3), 686–696. https://doi.org/10.1016/j.ejor.2014.09.041

Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research, 2*(6), 429–444. https://doi.org/10.1016/0377-2217(78)90138-8

Chen, Y., Tsionas, M. G., & Zelenyuk, V. (2021). LASSO+DEA for small and big wide data. *Omega, 102*, Article 102419. https://doi.org/10.1016/j.omega.2021.102419

Cherchye, L., Moesen, W., & Van Puyenbroeck, T. (2004). Legitimately diverse, yet comparable: On synthesizing social inclusion performance in the EU. *JCMS Journal of Common Market Studies, 42*(5), 919–955. https://doi.org/10.1111/j.0021-9886.2004.00535.x

Cherchye, L., Lovell, C. K., Moesen, W., & Van Puyenbroeck, T. (2007a). One market, one number? a composite indicator assessment of EU internal market dynamics. *European Economic Review, 51*(3), 749–779. https://doi.org/10.1016/j.euroecorev.2006.03.011

Cherchye, L., Moesen, W., Rogge, N., & Van Puyenbroeck, T. (2007b). An introduction to 'benefit of the doubt' composite indicators. *Social Indicators Research, 82*, 111–145. https://doi.org/10.1007/s11205-006-9029-7

Cherchye, L., Moesen, W., Rogge, N., & Van Puyenbroeck, T. (2011). Constructing composite indicators with imprecise data: A proposal. *Expert Systems with Applications, 38*(9), 10940–10949. https://doi.org/10.1016/j.eswa.2011.02.136

Cooper, W. W., Seiford, L., & Tone, K. (2007). *Data envelopment analysis: A comprehensive text with models, applications, references and DEA-solver software*. New York: Springer.

De Witte, K., & Rogge, N. (2010). To publish or not to publish? on the aggregation and drivers of research performance. *Scientometrics, 85*(3), 657–680. https://doi.org/10.1007/s11192-010-0286-5

De Witte, K., & Rogge, N. (2011). Accounting for exogenous influences in performance evaluations of teachers. *Economics of Education Review, 30*(4), 641–653. https://doi.org/10.1016/j.econedurev.2011.02.002

Doyle, J., & Green, R. (1994). Efficiency and cross-efficiency in DEA: Derivations, meanings and uses. *Journal of the Operational Research Society, 45*(5), 567–578. https://doi.org/10.1057/jors.1994.84

Dyson, R. G., Allen, R., Camanho, A. S., Podinovski, V., Sarrico, C. S., & Shale, E. A. (2001). Pitfalls and protocols in DEA. *European Journal of Operational Research, 132*(2), 245–259. https://doi.org/10.1016/S0377-2217(00)00149-1

EUROSTAT (2023). Sustainable Development Indicators database. https://ec.europa.eu/eurostat/web/sdi/database.

Ferreira, D. C., Caldas, P., Varela, M., & Marques, R. C. (2023). A geometric aggregation of performance indicators considering regulatory constraints: An application to the urban solid waste management. *Expert Systems with Applications, 218*, Article 119540. https://doi.org/10.1016/j.eswa.2023.119540

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. https://doi.org/10.1016/S0377-2217(97)84108-2

Fusco, E. (2015). Enhancing non-compensatory composite indicators: A directional proposal. *European Journal of Operational Research, 242*, 620–630. https://doi.org/10.1016/j.ejor.2014.10.017

Fusco, E. (2023). Potential improvements approach in composite indicators construction: The Multi-directional Benefit of the Doubt model. *Socio-Economic Planning Sciences, 85*, Article 101447. https://doi.org/10.1016/j.seps.2022.101447

Fusco, E., Vidoli, F., & Rogge, N. (2020). Spatial directional robust Benefit of the Doubt approach in presence of undesirable output: An application to italian waste sector. *Omega, 94*, Article 102053. https://doi.org/10.1016/j.omega.2019.03.011

Fusco, E., Vidoli, F., & Sahoo, B. K. (2018). Spatial heterogeneity in composite indicator: A methodological proposal. *Omega, 77*, 1–14. https://doi.org/10.1016/j.omega.2017.04.007

Ghasemi, M. R., Ignatius, J., & Rezaee, B. (2019). Improving discriminating power in data envelopment models based on deviation variables framework. *European Journal of Operational Research, 278*(2), 442–447. https://doi.org/10.1016/j.ejor.2018.08.046

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

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.

Homburg, C. (2001). Using data envelopment analysis to benchmark activities. *International Journal of Production Economics, 73*(1), 51–58. https://doi.org/10.1016/S0925-5273(01)00194-3

Jablonsky, J. (2007). Measuring the efficiency of production units by AHP models. *Mathematical and Computer Modelling, 46*(7–8), 1091–1098. https://doi.org/10.1016/j.mcm.2007.03.007

Jenkins, L., & Anderson, M. (2003). A multivariate statistical approach to reducing the number of variables in Data Envelopment Analysis. *European Journal of Operational Research, 147*(1), 51–61. https://doi.org/10.1016/S0377-2217(02)00243-6

Kapelko, M., Ortiz, L., & Aparicio, J. (2024). Comparing groups of units through composite indicators in a non-convex approach: Corporate social responsibility for the food and beverage manufacturing industry. *Annals of Operations Research, forthcoming.*. https://doi.org/10.1007/s10479-024-06139-6

Karagiannis, G. (2017). On aggregate composite indicators. *Journal of the Operational Research Society, 68*(7), 741–746. https://doi.org/10.1057/jors.2015.81

Karagiannis, R., & Karagiannis, G. (2018). Intra-and inter-group composite indicators using the BoD model. *Socio-Economic Planning Sciences, 61*, 44–51. https://doi.org/10.1016/j.seps.2017.01.002

Koronakos, G., Kritikos, M., & Sotiros, D. (2024). A common weights multiplicative aggregation approach for composite indicators: The case of Global City Competitiveness Index. *Expert Systems with Applications, 242*, Article 122543. https://doi.org/10.1016/j.eswa.2023.122543

Kuosmanen, T., & Johnson, A. L. (2010). Data Envelopment Analysis as nonparametric least-squares regression. *Operations Research, 58*(1), 149–160. https://doi.org/10.1287/opre.1090.0722

Lahouel, B. B., Zaied, Y. B., Song, Y., & Yang, G. L. (2021). Corporate social performance and financial performance relationship: A data envelopment analysis approach without explicit input. *Finance Research Letters, 39*, Article 101656. https://doi.org/10.1016/j.frl.2020.101656

Lee, C.-Y., & Cai, J.-Y. (2020). Lasso variable selection in Data Envelopment Analysis with small datasets. *Omega, 91*, Article 102019. https://doi.org/10.1016/j.omega.2018.12.008

Li, Q. (1996). Nonparametric testing of closeness between two unknown distribution functions. *Econometric Reviews, 15*(3), 261–274. https://doi.org/10.1080/07474939608800355

Limleamthong, P., & Guillén-Gosálbez, G. (2018). Mixed-integer programming approach for dimensionality reduction in Data Envelopment Analysis: Application to the sustainability assessment of technologies and solvents. *Industrial & Engineering Chemistry Research, 57*(30), 9866–9878. https://doi.org/10.1021/acs.iecr.7b05284

Lovell, C. A. K., & Pastor, J. T. (1999). Radial DEA models without inputs or without outputs. *European Journal of Operational Research, 118*, 46–51. https://doi.org/10.1016/S0377-2217(98)00338-5

Mergoni, A. (2022). Package 'rcDEA', release October 14, 2022. Version 1.0. https://cran.r-project.org/web/packages/rcDEA/rcDEA.pdf.

Nataraja, N. R., & Johnson, A. L. (2011). Guidelines for using variable selection techniques in Data Envelopment Analysis. *European Journal of Operational Research, 215*(3), 662–669. https://doi.org/10.1016/j.ejor.2011.06.045

Nunamaker, T. R. (1985). Using data envelopment analysis to measure the efficiency of non-profit organizations: A critical evaluation. *Managerial and Decision Economics, 6*(1), 50–58. https://doi.org/10.1002/mde.4090060109

Oliveira, R., Zanella, A., & Camanho, A. S. (2019). The assessment of corporate social responsibility: The construction of an industry ranking and identification of potential for improvement. *European Journal of Operational Research, 278*(2), 498–513. https://doi.org/10.1016/j.ejor.2018.11.042

Pastor, J. T., Aparicio, J., & Zofío, J. L. (2022). *Benchmarking economic efficiency*. Cham: Springer.

Pastor, J. T., Ruiz, J. L., & Sirvent, I. (2002). A statistical test for nested radial DEA models. *Operations Research, 50*(4), 728–735. https://doi.org/10.1287/opre.50.4.728.2866

Pereira, M. A., Camanho, A. S., Marques, R. C., & Figueira, J. R. (2021). The convergence of the world health organization member states regarding the united nations' sustainable development goal 'good health and well-being'. *Omega, 104*, Article 102495. https://doi.org/10.1016/j.omega.2021.102495

Peyrache, A., Rose, C., & Sicilia, G. (2020). Variable selection in Data Envelopment Analysis. *European Journal of Operational Research, 282*(2), 644–659. https://doi.org/10.1016/j.ejor.2019.09.028

Prieto, A. M., & Zofío, J. L. (2001). Evaluating effectiveness in public provision of infrastructure and equipment: The case of spanish municipalities. *Journal of Productivity Analysis, 15*, 41–58. https://doi.org/10.1023/A:1026595807015

Raab, R. L., & Lichty, R. W. (2002). Identifying subareas that comprise a greater metropolitan area: The criterion of county relative efficiency. *Journal of Regional Science, 42*(3), 579–594. https://doi.org/10.1111/1467-9787.00273

Rogge, N. (2018a). Composite indicators as generalized benefit-of-the-doubt weighted averages. *European Journal of Operational Research, 267*(1), 381–392. https://doi.org/10.1016/j.ejor.2017.11.048

Rogge, N. (2018b). On aggregating benefit of the doubt composite indicators. *European Journal of Operational Research, 264*(1), 364–369. https://doi.org/10.1016/j.ejor.2017.06.035

Ruggiero, J. (2005). Impact assessment of input omission on DEA. *International Journal of Information Technology & Decision Making, 4*(03), 359–368. https://doi.org/10.1142/S021962200500160X

Sahoo, B. K., & Acharya, D. (2012). Constructing macroeconomic performance index of Indian states using DEA. *Journal of Economic Studies, 39*(1), 63–83. https://doi.org/10.1108/01443581211192116

Simar, L., & Wilson, P. W. (2000). A general methodology for bootstrapping in non-parametric frontier models. *Journal of Applied Statistics, 27*(6), 779–802. https://doi.org/10.1080/02664760050081951

Simar, L., & Zelenyuk, V. (2006). On testing equality of distributions of technical efficiency scores. *Econometric Reviews, 25*(4), 497–522. https://doi.org/10.1080/07474930600972582

Torgersen, A. M., Førsund, F. R., & Kittelsen, S. A. (1996). Slack-adjusted efficiency measures and ranking of efficient units. *Journal of Productivity Analysis, 7*, 379–398. https://doi.org/10.1007/BF00162048

United Nations (2015). Transforming Our World: The 2030 Agenda for Sustainable Development. https://sdgs.un.org/2030agenda.

Van Puyenbroeck, T. (2018). On the output orientation of the Benefit-of-the-Doubt-Model. *Social Indicators Research, 139*, 415–431. https://doi.org/10.1007/s11205-017-1734-x

Van Puyenbroeck, T., & Rogge, N. (2020). Comparing regional human development using global frontier difference indices. *Socio-Economic Planning Sciences, 70*, Article 100663. https://doi.org/10.1016/j.seps.2018.10.014

Vidoli, F., Fusco, E., & Mazziotta, C. (2015). Non-compensability in composite indicators: A robust directional frontier method. *Social Indicators Research, 122*, 635–652. https://doi.org/10.1007/s11205-014-0710-y

Vidoli, F., Fusco, E., Pignataro, G., & Guccio, C. (2024). Multi-directional Robust Benefit of the Doubt model: An application to the measurement of the quality of acute care services in OECD countries. *Socio-Economic Planning Sciences, 93*, Article 101877. https://doi.org/10.1016/j.seps.2024.101877

Walheer, B. (2024). A sequential benefit-of-the-doubt composite indicator. *European Journal of Operational Research, 316*(1), 228–239. https://doi.org/10.1016/j.ejor.2024.01.029

Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software, 6*(60), 3021. https://doi.org/10.21105/joss.03021

Wilson, P. W. (2018). Dimension reduction in nonparametric models of production. *European Journal of Operational Research, 267*(1), 349–367. https://doi.org/10.1016/j.ejor.2017.11.020

Yamada, Y. (1994). An inefficiency measurement method for management systems. *Journal of the Operations Research Society of Japan, 37*, 158–168. https://doi.org/10.15807/jorsj.37.158

Zanella, A., Camanho, A. S., & Dias, T. G. (2015). Undesirable outputs and weighting schemes in composite indicators based on data envelopment analysis. *European Journal of Operational Research, 245*, 517–530. https://doi.org/10.1016/j.ejor.2015.03.036