Socio-Economic Planning Sciences 106 (2026) 102528

Contents lists available at ScienceDirect

Socio-Economic Planning Sciences

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

# Utility effectiveness, composite indicators and 'benefit of the doubt' models: How important are preferences when evaluating the performance of innovation systems?

José Luis Zofío [a,b,*], Juan Aparicio [c,d], Javier Barbero [a], Jon Mikel Zabala-Iturriagagoitia [e,f,g]

[a] Department of Economics, Universidad Autónoma de Madrid, Madrid, Spain
[b] Rotterdam School of Management, Erasmus University, Rotterdam, the Netherlands
[c] Center of Operations Research (CIO), University Miguel Hernandez of Elche (UMH), Elche, E-03202, Spain
[d] Joint Research Unit, valgrAI - Valencian Graduate School and Research Network of Artificial Intelligence, Camino de Vera s/n, Valencia, 46022, Spain
[e] Deusto Business School, University of Deusto, San Sebastián, Spain
[f] South Eastern University Norway, Kongsberg, Norway
[g] CIRCLE, Lund University, Lund, Sweden

ARTICLE INFO

*Keywords:*
Utility effectiveness
Composite indicators
Benefit-of-the-doubt evaluation
Innovation systems
Innovation performance

ABSTRACT

This study introduces a methodology that embeds 'Benefit of the Doubt' evaluations of performance of innovation systems, based on objective measures of technical effectiveness—that compare the values of a set of indicators across countries—into the literature of composite indicators, which rank countries relying on utility functions and whose aggregating structure and weights are subjectively chosen by the decision makers. Considering the general case of the constant elasticity of substitution utility function, whose power mean formulation nests popular specifications such as the weighted (linear) arithmetic mean, weighted geometric mean (Cobb-Douglas), and fixed-proportions (Leontief), we define the concept of utility effectiveness of an alternative as the ratio of its observed utility to the maximum utility observed across countries. We show that the utility loss resulting from an ineffective performance can be consistently decomposed into two mutually exclusive factors: first, grounded in the objective criterion of Pareto-dominance, the mentioned technical effectiveness is represented by a technical (radial) measure depending only on quantities, which gauges the distance to the utility frontier; second, a measure of allocative effectiveness, determining the additional utility loss resulting from an unbalanced composition of indicators given the preferences of the decision maker. Using data from the European Innovation Scoreboard, our results show high levels of utility ineffectiveness in Southern and Eastern European countries. On average, these countries' performance is half of the maximum utility reached by the best performing innovation systems. Objective technical ineffectiveness generally dominates over subjective allocative ineffectiveness (except for the fixed proportion specification), showing that the practical impact of choosing one utility specification over another, although sizable, is limited. Consequently, main performance deficits are often not due to substitutability or compensatory issues, but rather to real shortfalls in indicator levels, while preference-driven differences tend to be second-order and comparatively robust.

## 1. Introduction

Composite indicators (CIs) and scoreboards are widely used in real-world policymaking to condense multidimensional evidence into an intelligible message for benchmarking, peer learning, and public communication. In the European Union (EU), indices and scoreboards are regularly used in the preparatory stages of policy (e.g., impact assessments), for monitoring strategies and action plans, and policy dialogue [1]. A prominent example is the European Innovation Scoreboard (EIS), which provides an annual comparative assessment of Member States' research and innovation performance, highlights relative strengths and weaknesses, and helps governments identify areas where

* Corresponding author. Department of Economics, Universidad Autónoma de Madrid, Madrid, Spain.
*E-mail addresses:* jose.zofio@uam.es, jzofio@rsm.nl (J.L. Zofío).

https://doi.org/10.1016/j.seps.2026.102528
Received 27 October 2025; Received in revised form 2 June 2026; Accepted 3 June 2026
Available online 11 June 2026
0038-0121/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

efforts should be concentrated [2]. This institutional role clarifies the main field of application of CIs: they are primarily tools for communication and performance monitoring, rather than scientific measurement instruments in the strict sense.

At the same time, educated decision-making requires CI methodologies that are consistent and transparent, otherwise, almost any reallocation of resources could be justified ex post. Early studies by Freudenberg [3] and Bandura [4] provide critical assessments and document the proliferation of CIs. On the one hand, CIs are valuable for their ability to synthesize large datasets into comprehensible formats, supporting communication and political discourse thanks to their simplicity [5]. On the other, construction choices (indicator selection, normalization, weighting, aggregation) raise methodological and governance challenges that may foster misinterpretation, strategic use, and disputes over credibility [6,7]. Reviews of the methodological framework emphasize that one of the most controversial steps concerns the choice of the aggregating function and weights [8–11].

These controversies are not purely technical; they rather reflect the institutional constraints under which CIs operate. Policy institutions often favor simple and stable constructions (e.g., linear aggregation with equal weights) because they are easy to explain, comparable over time, and perceived as less vulnerable to accusations of 'rigging'. Yet, the choice of an aggregating function is not neutral because it encodes assumptions about compensability/substitutability among indicators [12]. In economic terms, the aggregating function represents the preference structure—or more precisely, the utility—of a decision maker (DM). In this context, a DM is the agent whose preferences determine the aggregation rule, the degree of substitutability among indicators, and their relative importance through weights. Typical DMs include international organizations, public institutions, or research bodies that design and use composite indicators for benchmarking and policy communication. In our empirical application the DM is the European Commission (EC), responsible for the EIS. By selecting the indicators, adopting an equal-weighting scheme, and choosing a linear aggregation logic, the EC implicitly defines the utility structure used to evaluate national innovation systems, while countries are the alternatives under evaluation.

As discussed in the methodological section, there are multiple ways to aggregate indicators and evaluate performance. In practice, many CI producers—including the EC—favor the linear arithmetic mean, which assumes perfect substitutability across indicators (mutual preferential independence), so that trade-offs among any subset of attributes do not depend on the remaining subset. This strong compensation assumption has been criticized as unrealistic, and alternative specifications that allow imperfect substitutability (e.g., geometric means) or no substitutability (fixed-proportions) are often advocated [13]. For example, if positive amounts of each variable are required to generate any utility (i.e., zeros are excluded), geometric mean or fixed-proportions formulations provide more adequate representations because they require all indicators to be present in some quantity. By contrast, when zeros occur, linear aggregation may be the only operational option.

Besides the functional form, establishing relative weights is fundamental. Given a prespecified utility function, a weight determines each indicator's contribution to aggregate utility; weights define trade-offs between indicators and therefore, their relative importance within the chosen utility structure. Consequently, selecting a utility function cannot be separated from selecting weights. The function captures the structure of preferences, while the weights determine their specific valuation. Because both choices convey the DM's preferences, CI construction is inherently supervised and may be contested by evaluated units. In supranational benchmarking, for instance, countries may disagree with the common weights embedded in a CI. However, allowing each country to select its own weights within the same functional form introduces an additional layer of discretionality.

A well-known response is to adopt an unsupervised approach in which weights are endogenously determined to give each evaluated unit the 'benefit of the doubt'. This is precisely what Data Envelopment Analysis (DEA) techniques and, more specifically, the so-called 'Benefit-of-the-Doubt' (BoD) model do [8,9,14]. BoD constructs an empirical frontier and evaluates each decision making unit by choosing the most favorable multipliers (weights) consistent with the data, thereby providing an objective, quantity-based benchmark. Based on mathematical programming, the DEA-BoD dual model evaluates the utility of an alternative (also known as decision making unit) by searching for a supporting hyperplane that allows it to reach the maximum value of the CI, which corresponds to a normalized effectiveness score. This hyperplane is defined as a convex combination of the observed variables, implying that the aggregating function corresponds to the linear weighted mean. In the primal formulation, adopted in this study, DEA-BoD methods construct a polyhedral of piecewise segments corresponding to linear combinations of observed alternatives that, enveloping the data, define the attainable set. The evaluation is made in relative terms (i.e., comparing all alternatives against each other), and the flexibility in the value of the weights (in the dual formulation) and the multipliers (in the primal formulation) implies that non-dominated alternatives are deemed effective (i.e., they belong to a segment known as effective frontier). In this case, under the identified optimal weights and multipliers, the alternative reaches the maximum possible value of the CI (utility), laying on the utility frontier, with a normalized score of one. Numerous extensions of the original BoD radial framework have been proposed, including slack-based formulations [15], approaches accommodating imprecise or interval data [16], models that explicitly incorporate weight restrictions across indicators [17], undesirable indicators [18], conditional and robust approaches accounting for external factors [19,20], non-compensatory CIs [21], spatially robust models [22,23], intra- and inter-group approaches controlling for heterogeneity [24,25], consistency in aggregation methods [26–29], directional and goal-programming variants [30–32], and recent robust and nonconvex formulations [33–37] including 'Ensemble-DEA' for robust rankings and benchmarking [38].

Therefore, DEA-BoD can be considered as valid alternative to construct CIs that, relying on a specific aggregating function (the linear weighted mean), allows for an endogenous and objective (flexible) selection of weights; one that ensures that the different alternatives are deemed as effective as possible in their own, most favorable, terms (dimensions). However, full flexibility also creates practical interpretability problems in policy settings. BoD may assign zero weights to indicators that stakeholders regard as essential, and it may label as effective any alternative that excels in a single dimension irrespective of weaknesses elsewhere. This can yield implausible 'virtual' CIs and rankings that misrepresent best-practice performance from a systems perspective. The issue is aggravated when the number of observations is limited relative to the number of indicators (curse of dimensionality), which reduces discriminatory power. In our empirical application we address this using 'Ensemble-DEA' [38]. These limitations are not minor in political environments where CIs must be credible, interpretable, and resistant to strategic narratives.

The above discussion shows that neither subjective and supervised methods like the CI methodology, nor objective and unsupervised methods like DEA-BoD are fully satisfactory, since both have advantages and disadvantages. The literature has proposed bridges between supervised CI approaches and unsupervised BoD by incorporating preferences into DEA-BoD. Two prominent remedies are (i) weight restrictions—via ordinal ranks [8], assurance regions [39], pie shares [9,40] or absolute bounds [41]—, and (ii) methods that identify most-preferred solution(s) and restrict attention to frontier facets containing them, using approaches ranging from Pareto-race [42] and TOPSIS-based ideals [43] to balance/'uniformity' heuristics. A limitation is that these approaches mainly modify the shape of the frontier but typically do not incorporate the CI methodology through an explicit utility function describing the DM's substitutability and weighting assumptions.

In this study we propose a third and complementary bridge. Our novel framework combines utility theory with DEA-BoD.[1] Utility theory specifies an explicit aggregating function and weights representing the DM's preference structure and identifies the Most Preferred Alternative (MPA) by solving a utility maximization problem over the same attainable set. We then define overall utility effectiveness as the ratio between the observed utility of an alternative and the utility of the MPA. As this evaluation is intrinsically dependent on subjective preferences, the resulting outcome may be challenged and ultimately unaccepted by stakeholders and interested parties. To qualify and interpret such contestation, we embed the objective BoD evaluation and define utility-technical effectiveness as the ratio between observed utility and the utility of the BoD benchmark on the frontier, which corresponds to a preference-free, quantity-driven measure grounded in the undisputed concept of Pareto-effectiveness. That is, an alternative is effective, defining the utility frontier, if it is not possible to find other alternative that dominates it, namely, that has a larger value in at least one variable and no less in the other variables. We deem this differential in utility effectiveness technical because it depends solely on quantities. However, the difference in the quantities of the variables is measured in terms of—or translated into—their corresponding CIs through the chosen aggregating utility function.

Crucially, we show that overall utility effectiveness decomposes into (i) utility-technical effectiveness, which coincides with (the inverse of) DEA–BoD effectiveness and is independent of the utility function, its substitutability parameter and weights, and (ii) utility-allocative effectiveness, a factor capturing the preference-driven gap due to the composition (mix/balance) of indicators relative to the DM's trade-offs. Because the difference between the overall and technical utility effectiveness corresponds to the solution of the utility maximization problem under the structure of preferences of the DM, and the solution of the DEA-BoD model is the result of an unsupervised maximization process, any difference can be attributed to the discretional authority of the DM. This decomposition provides the institutional improvement sought by policymakers: it separates the objective part of performance shortfalls (insufficient indicator levels relative to best practice) from the subjective part (the extent to which performance depends on contested aggregation assumptions). When technical ineffectiveness dominates, performance deficits are primarily real shortfalls in indicator quantities; when allocative ineffectiveness is sizeable, rebalancing the composition of indicators becomes central.

We illustrate the approach by evaluating EU innovation systems using EIS data. The EIS is explicitly intended as a comparative monitoring tool for Member States and is used to identify strengths, weaknesses, and priority areas for policy [2]. Our empirical analysis compares results under three representative degrees of substitutability (or compensability, as it is referred to in the CI literature) and quantifies whether debates about aggregation and weights materially affect conclusions once the objective technical component is separated from the subjective allocative component.

This article develops as follows. In Section 2 we present the theory to measure subjective utility effectiveness. We develop our model considering choice theory that assumes that DMs exhibit: a) a coherent preference structure over a set of variables, characterized by a utility (aggregating) function with corresponding weights, and b) a rational behavior by choosing the one that maximizes it subject to the attainable set of alternatives. In this section we consider as reference the constant elasticity of substitution utility function, which nests the most popular specifications used in constructing CIs, with the above degrees of substitutability: arithmetic mean (linear), geometric mean (Cobb-Douglas), and fixed proportions (Leontief). We close this section defining the concept of overall utility effectiveness and providing an example that illustrates it. In Section 3 we recall the DEA-BoD methods, define the concept of technical utility effectiveness and relate it to the overall utility effectiveness. We also show the decomposition of utility effectiveness into technical (quantities) based on the concept of Pareto-effectiveness and the residual allocative (weights) utility effectiveness that can be attributed to alternative sets of weights (those favoring the evaluated alternative under DEA-BoD versus those chosen by the DM). Section 4 is devoted to our application. Here we report the results on utility effectiveness comparing the above aggregating functions with their corresponding degrees of substitutability: linear-perfect substitutes, Cobb–Douglas-imperfect substitutes, and Leontief-no substitutability, as well as its decomposition into technical (DEA-BoD) and allocative effectiveness. In the calculations of technical effectiveness and to prevent the 'curse of dimensionality', we adopt 'Ensemble-DEA' and identify the benchmark peers for each country. Finally, Section 5 concludes by discussing its contributions and identifying the potential avenues for further research.

## 2. Utility maximization and benchmarking: utility effectiveness

### 2.1. Utility functions and substitutability

Let us consider a DM who wants to determine the relative performance of $k = 1, \ldots, M$ observed alternatives (individuals, firms, organizations, countries) according to a utility function (or CI) representing her preferences over a set of non-negative $i = 1, \ldots, N$ variables (criteria, attributes, indicators). Following choice theory (e.g., Ref. [13]), we assume that the DM has a rational, continuous, and locally nonsatiated preference relation on the attainment set of variables, represented by a utility function that is assumed to be continuous. Additional properties imposing regularity conditions such as differentiability, monotonicity and quasiconcavity may be considered also for analytical convenience. These properties ensure that the indifference sets are convex and exhibit a decreasing marginal rate of substitution between variables. In this study we consider that the utility function of the DM is represented by a Constant Elasticity of Substitution (CES) aggregator [46]:

$$U(I) = \left[w_1 I_1^{\rho} + w_2 I_2^{\rho} + \ldots + w_N I_N^{\rho}\right]^{1/\rho}, \tag{1}$$

where $\rho \in [-\infty, 1]$ is the substitution parameter, $w_i \in [0, 1]$, $\sum_i w_i = 1$, $i = 1, \ldots, N$, are preference parameters weighting the variables, and $I_i$, $i = 1, \ldots, N$, are the observed quantities of the variables, or, considering our empirical application, indicators. Relevant for this study, the CES utility function (1) is homogenous of degree one, implying that if we increase the quantity of all indicators in a given proportion, utility increases in the same proportion (nevertheless, it can be modified to accommodate other degrees of homogeneity). Although other utility functions have been proposed (e.g., the quasi-linear utility function), the CES specification is chosen because it corresponds to a generalized weighted average (power mean) that nests the three most popular functions as special cases, capturing the whole range of elasticity of substitution between infinitum and zero. Each of these functions characterize a different degree of substitutability between variables, while keeping the utility level constant. Let us define the elasticity of

[1] The methodology is inspired by Farrell's [44] decomposition of economic (revenue) inefficiency. Revenue efficiency compares observed revenue with maximum feasible revenue (being the revenue function a linear weighted mean), which can be decomposed into a technical component (distance to the frontier) and an allocative component (suboptimal output mix given market prices). Once projected onto the frontier, a firm may still fail to maximize revenue because its output mix is not aligned with prices; the remaining gap is allocative inefficiency [45]. This logic shows how objective frontier benchmarking and preference weight/price structures can be combined rather than substituted.

substitution among indicators as $\sigma = 1/(1-\rho)$, representing the percentage change in the ratio of two variables divided by the percentage change in their Marginal Rate of Substitution (MRS) while keeping the utility level constant, then[2]

- If $\rho \to 1$, $U^{LR}(I) = \sum_{i=1}^{N} w_i I_i$ is linear (LR), and the variables are perfect substitutes: $\sigma \to \infty$. Utility is calculated as the weighted arithmetic mean of the variables.
- If $\rho \to 0$, then $U^{CD}(I) = \prod_{i=1}^{N} I_i^{w_i}$ corresponds to the Cobb-Douglas (CD) specification: $\sigma \to 1$. Utility is calculated as the weighted geometric mean of the variables.
- If $\rho \to -\infty$, then $U^{LT}(I) = \min\{I_i, i = 1, \ldots, N\}$ corresponds to the Leontief (LT) specification and the variables are perfect complements: $\sigma \to 0$, i.e., there is no substitutability among them.

Table 1 summarizes the above specifications and features.

From a quantitative perspective, the following relationship between the three specifications holds:

**Proposition 1**. *Given a set of variables $I_i$, $i$=1, …,$N$, with $I_i \geq 0\ \forall\ i$, and associated weights $w_i \in (0,1]$, $\sum_i w_i = 1$, $i$=1, …,$N$, the following inequalities are observed $U^{LR}(I) \geq U^{CD}(I) \geq U^{LT}(I)$.*

*Proof. First, as for the inequality $U^{LR}(I) \geq U^{CD}(I)$, since $U^{LR}(I) = \sum_i^N w_i I_i$ is a weighted arithmetic mean and $U^{CD}(I) = \prod_{i=1}^{N} I_i^{w_i}$ is a weighted geometric mean, Aldaz [47] proves the following relationship: $\sum_i^N w_i I_i \geq \prod_{i=1}^{N} I_i^{w_i} + \sum_{i=1}^{N} w_i \left(I_i^{0.5} - \sum_{i=1}^{N} w_i I_i^{0.5}\right)^2$, where the last term is the weighted variance of the vector $I^{0.5} = (I_1^{0.5}, \ldots, I_N^{0.5})$. Therefore, given that $\sum_{i=1}^{N} w_i \left(I_i^{0.5} - \sum_{i=1}^{N} w_i I_i^{0.5}\right)^2 \geq 0$, $U^{LR}(I) \geq U^{CD}(I)$—the equality is verified when $I_i = I_j, \forall i,j$. Second, as for the inequality $U^{CD}(I) \geq U^{LT}(I)$, we assume without loss of generality that $U^{LT}(I) = \min\{I_i, i = 1, \ldots, N\} = I_1$, then $I_i \geq I_1$ $\forall i$. Since $w_i \geq 0$, raising both sides to the power $w_i$ preserves the inequality: $I_1^{w_i} \geq I_1^{w_i}$, $\forall i$. Multiplying over $i$=1, …,$N$ yields $\prod_{i=1}^{N} I_i^{w_i} \geq \prod_{i=1}^{N} I_1^{w_i} = I_1^{w_1+\ldots+w_N} = I_1$, i.e. $U^{CD}(I) \geq U^{LT}(I) U^{CD}(I)$—again, the equality is verified when $I_i = I_j, \forall i,j$.*

Within the CES utility framework, the linear and Leontief specifications represent the extreme cases of substitutability among variables, respectively corresponding to the fully compensatory and non-compensatory approaches discussed in the literature [12]. Compensability in the linear utility function implies that variables can be fully substituted. For example, in an index used for ranking innovation systems, increasing the R&D expenditure in the business sector (BERD) can always substitute a reduction in the number of Employed ICT specialists (EICTS). Contrarily, under the Leontief utility function, indicators are non-compensatory and cannot be substituted. Although the measurement of satisfaction in terms of utility normally refers to a subject-dependent, ordinal concept, the definition and numerical calculation of CIs represent an area where utility theory provides guidance. The reason is that the preferences of a DM are clearly spelled out and commonly applied to a set of alternatives.

### 2.2. Discussion of the alternative utility specifications

The linear utility function $U^{LR}(I)$ is the most popular among DMs as it represents the usual aggregation rule applied to a set of indicators (OECD, 2003, p.5). The reason is that it allows to rank observations considering the weighted arithmetic mean of the indicators and without restricting the substitutability among them, thereby accommodating any value observed across the alternatives to be evaluated, including zeros. The Cobb-Douglas specification $U^{CD}(I)$ corresponds to the weighted geometric mean and the degree of substitutability is lower, while utility tends to zero as any of the indicators approaches zero. This function is by far the most popular one in economics because it is consistent with standard assumptions resulting in a decreasing MRS among variables (e.g., marginal utilities are decreasing in the variables—amounts consumed of the goods). However, the geometric mean penalizes observations whose indicators have low values and large variability (i.e., unbalanced values with indicators close to zero), which exacerbates differences among observations in the rankings, penalizing the worst performers.[3] Contrarily, the Leontief specification $U^{LF}(I)$ has never been considered in benchmarking exercises until now because, by precluding any substitutability, it requires DMs to explicitly define a specific set of fixed proportions among indicators (e.g., BERD/EICTS). In this function, the indicator with the smallest value represents a bottleneck and, therefore, this specification penalizes poor performance the most. In this regard, the more strict the attitude of the DM in terms of substitutability, in the sense that the focus is more on the indicators in which the alternatives perform poorly, without the possibility of compensating this deficit with those in which it performs better, the more negative the value for the substitution parameter $\rho$ should be (see Proposition 1 below). The Leontief case ($\rho = -\infty$) represents the extreme case as the utility corresponds to the minimum value observed across indicators.

Relevant to the determination of the substitutability among indicators is the choice of weights $w_i$. Different options are available. Some are derived from statistical and optimization models, such as Principal Components Analysis, Factor analysis, Benefit of the Doubt, and unobserved components models, or from participatory methods like budget allocation processes, analytic hierarchy processes and conjoint analysis [6]. However, most DMs rely on equal weighting, $w_i = 1/N$, $i = 1, \ldots, N$. In the case of the linear utility function this choice makes its value equivalent to the arithmetic average of the indicators: $U^{LR}(I) = \sum_i I_i/N$. In the case of the Cobb-Douglas specification the value corresponds to the geometric mean: $U^{CD}(I) = \left(\prod_i I_i\right)^{1/N} = \sqrt[N]{\prod_i I_i}$.

When choosing equal weights, indicators can be substituted in equal absolute values under the linear specification and in equal percentage values under the Cobb-Douglas specification (e.g., a 5% increase in BERD compensates a 5% reduction in EICTS), keeping the value of the utility constant. Moreover, focusing on the MRS, while the substitutability of indicators is constant under the linear specification, regardless the relative values of the indicators (high or low), it is decreasing for the Cobb-Douglas formulation. In terms of the indifference curves of the utility functions they are linear and convex, respectively.[4] The literature has criticized the simplicity of the linear utility function [8], but its specification, equivalent to the arithmetic mean, makes it more understandable to stakeholders than the geometric mean (Cobb-Douglas) or fixed proportions (Leontief) formulations. We share the criticism to the linear function because it is unrealistic to assume that indicators are fully compensable regardless of their relative abundance or scarcity. Using the previous example, increasing the BERD, can always substitute

[2] Proofs of these results normally resort to the fact that the CES function has the form of a generalized weighted (power) mean or applying L'Hôpital's rule to the differential of the CES function ([27], Appendix).

[3] For an interpretation of the utility functions as CIs constructed with penalizing power means, see Mariani et al. [48].

[4] This implies that while the marginal utility of increasing an indicator is constant under the linear specification, it is decreasing in the Cobb-Douglas specification. Hence, in benchmarking exercises, alternatives with lower values would tend to favor geometric aggregation over linear aggregation. The reason is that the marginal utility derived from an increase in a low valued indicator, ceteris paribus the remaining indicators, is greater under geometric aggregation than under the linear aggregation. Therefore, a country would be more incentivized to focus on indicators with low values if geometric aggregation were employed, as this approach would enhance its prospects of improving its ranking position at a greater pace.

**Table 1**
Representative cases of the CES utility function used in composite indicators.

| Utility | Substitution parameter: $\rho$ | Functional specification | Substitutab. (MRS) | Elasticity of substitution: $\sigma$ | Aggregating function |
|---|---|---|---|---|---|
| Linear (LR) | $\rho \to 1$ | $U^{LR}(I) = \sum_{i=1}^{N} w_i I_i$ | Perfect and constant | $\sigma \to \infty$ | Weighted arithmetic mean |
| Cobb-Doug. (CD) | $\rho \to 0$ | $U^{CD}(I) = \prod_{i=1}^{N} I_i^{w_i}$ | Variable and decreasing | $\sigma \to 1$ | Weighted geometric mean |
| Leontief (LT) | $\rho \to -\infty$ | $U^{LT}(I) = \min_i \{I_i\}$ | None | $\sigma \to 0$ | Weighted fixed proportions |

Note: Beyond these three representative cases bounding the elasticity of substitution, the CES utility function is a generalized weighted mean that produces a different aggregator for $\infty \geq \rho \geq -\infty$. The lower (higher) the value of $\rho$ the higher (lower) penalization for relatively poor performances on one or more variables.

a reduction in the EICTS. However, this substitutability must necessarily decrease as compensating equal reductions in the number of specialists becomes increasingly harder because they become so scarce that, eventually, it is not possible to have a functioning innovation system without their presence. On the contrary, with the linear specification, the same utility level can be achieved by constantly substituting one for another until only one of them is fully used and the other is zero.[5] Under the Leontief specification utility is independent of the weights, and the implicit proportionality is 1:1; i.e., the excess difference in any indicator whose value is greater than the minimum does not increase the utility. This is consistent with a preference scheme where all weights are equal, i.e., $w_i = 1/N$, $i = 1, \ldots, N$. As we demonstrate in what follows, the weights provide the necessary information to explicitly characterize the fixed proportions among indicators required by the DM. Moreover, from a policy perspective, matching the fixed proportions of the best performing alternatives may be a desirable goal for the DM to provide effectiveness improvement recommendations.

Summing up, we showed that by considering the CES utility function (1) we can easily model and interpret the most popular specifications and determine the relative performance of the alternatives obtained under each substitutability assumption, providing relevant and complementary information to DMs. This will be illustrated in the empirical section evaluating the performance of innovation systems where we determine the efforts that ineffective countries need to undertake to improve and eventually reach the utility levels of leading countries, and how the different specifications of the utility functions drive these changes.

### 2.3. Attainable set and its effective subsets

Assuming a utility function, the problem is to determine what alternative maximizes the utility of the DM, thereby becoming the MPA among the set of observations. For this purpose, let us define the attainable set, $T \subset \mathbb{R}_+^N$, considering the values of the indicators observed across the alternatives, in the following way:

$$T = \left\{ I \in \mathbb{R}_+^N : \sum_{k=1}^{K} \lambda_k I_{ki} \geq I_i, i = 1, \ldots, N; \ \sum_{k=1}^{K} \lambda_k = 1, \lambda_k \geq 0, k = 1, \ldots, M \right\}. \tag{2}$$

Here, the multipliers $\lambda_k$ are the intensity weights that, depending on their values, result in alternative linear combinations of the observed alternatives. They can be interpreted as 'percent contributions' given the constraint $\sum_{k=1}^{M} \lambda_k = 1$. As we show below, when searching for the alternative that maximizes the utility of the DM or the effectiveness score, these multipliers build the optimal benchmarks on the frontier for the alternative under evaluation.

The attainable set satisfies the following properties:

(P1) $T \neq \emptyset$;

(P2) $T$ is bounded $\forall I \in \mathbb{R}_+^N$;

(P3) $I \in T, I \geq I' \Rightarrow I' \in T$, i.e., indicators are freely disposable;

(P4) $T$ is a closed set;

(P5) $T$ is a convex set.

Fig. 1 illustrates the attainability set considering six observations and two indicators, whose values, following standard practice in the CI literature, are normalized to ensure that they are free of units of measurement and range in the unit interval, i.e., [0,1].[6] The attainable set is represented by the area below linear combinations of alternatives A, B and C, including the horizontal and vertical extensions, corresponding to property (P3). In this example, observations D and E lie inside the attainable set, and they will never be preferred by the DM.

It is worth noting the similarity between the attainable set $T$ and the Activity Analysis model proposed by Koopmans [50] in production theory. This model has been developed using DEA techniques starting with Farrell [44], Boles (1966) and Charnes et al. [51], evaluating the performance of decision making units. Readers familiar with this approach can relate the attainable set (2) to a production model without inputs [52] which is the base for the measurement of relative effectiveness (i.e., to what extent observations are capable of attaining specific goals), without considering the resources (efforts) employed in the process [53]. This same model is the base of the DEA-BoD approach for constructing CIs [9,54]. We resort to DEA methods to measure and decompose the utility effectiveness of the alternatives in the following sections.

A relevant concept for benchmarking and performance measurement is the frontier of the attainable set, represented by those observations

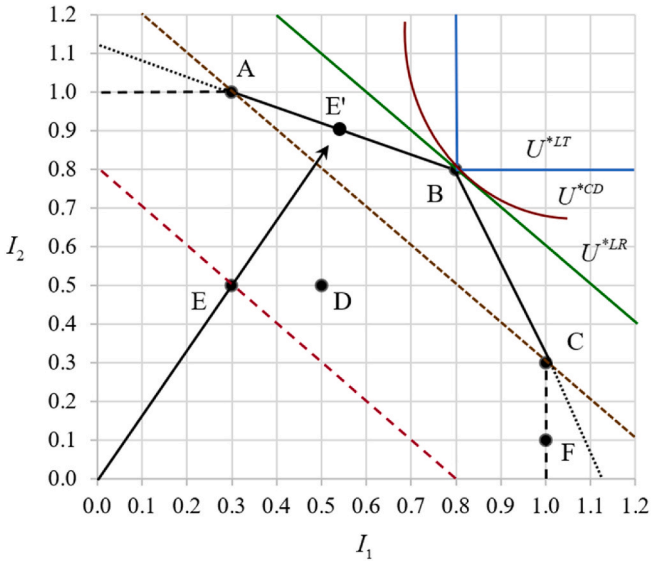



**Fig. 1.** An example illustrating the MPA for the CES function and utility effectiveness.

[5] This is the reason why the UNDP's Human Development Index switched from the arithmetic to the geometric mean, both with equal weights, in 2010, UNDP [49].

[6] See OECD [6], p.27) for a discussion on normalization methods.

that are not quantitatively dominated in the sense of Pareto (i.e., alternatives that lie on the boundary of $T$ and represent combinations of variables that cannot be improved upon a given dimension without worsening other dimensions). The frontier plays a key role in the evaluation of performance because it holds alternatives that are candidates to maximize the utility of the DM; that is, the MPA. Formally, the Pareto effective frontier corresponds to the following subset of $T$:

**Definition 1.** The strongly effective frontier of the attainability set $T$ is defined as

$$F^S(T) = \{I \in T : I' \geq I, I' \neq I \Rightarrow I' \notin T\} \tag{3}$$

In Fig. 1 the strongly effective frontier of the attainability set corresponds to the line segments ABC (i.e., observations A, B and C and the linear combinations between A and B and between B and C). In the benchmark analysis we consider all observations belonging to $F^S(T)$ as being strongly effective.

However, there is a weaker notion of effectiveness represented by alternatives that have the maximum value in one variable and therefore are not dominated in all dimensions. This is the case of observation F = $(I_{F1}, I_{F2}) = (1, 0.1)$ in Fig. 1, which shows the maximum value in the first indicator, but does not belong to $F^S(T)$ because it is dominated by observation C producing the same amount of $I_1$, but a larger quantity of $I_2$. Therefore, one can consider a weaker notion of effectiveness by recognizing these observations as part of a weaker notion of effective frontier.

**Definition 2.** The weakly effective frontier of the attainability set $T$ is defined as

$$F^W(T) = \{I \in T : I' > I \Rightarrow I' \notin T\} \tag{4}$$

In Fig. 1 the weakly effective subsets correspond to the boundary of the attainability set represented by the segments ABC (the strongly effective frontier) as well as their horizontal and vertical extensions (depicted as horizontal and vertical dashed lines). Therefore, $F^S(T) \subseteq F^W(T)$ and a strongly effective alternative is also weakly effective, but not vice versa. In general, the MPA will be strongly efficient but, as we show below, under specific utility specifications and weights (e.g., Leontief), the alternative maximizing the utility may be weakly efficient.

### 2.4. The utility maximization problem

The utility maximization problem (UMP) of the DM subject to the attainment set of alternatives (2) is formulated as follows:

$$\begin{aligned} &\underset{I,\lambda}{\text{Max}}\ U(I) = \left[w_1 I_1^{\rho} + w_2 I_2^{\rho} + \ldots + w_N I_N^{\rho}\right]^{1/\rho} \\ &\text{s.t. } T = \Bigg\{ I \in \mathbb{R}_+^N : \sum_{k=1}^{K} \lambda_k I_{ki} \geq I_i, i = 1, \ldots, N; \\ &\sum_{k=1}^{K} \lambda_k = 1, \lambda_k \geq 0, k = 1, \ldots, M \Bigg\}. \end{aligned} \tag{5}$$

In the UMP the DM chooses the combination of indicators in the attainable set $T$ to maximize her utility level, $U(I^*_{MPA})$. Given the polyhedral nature of the attainment set $T$, the preferred combination of indicators maximizing the utility, $I^*_{MPA}$, corresponds to a linear combination of the observed alternatives for which all $M$ restrictions are saturated (verified as equalities). Then the MPA can be recovered from the set of optimal intensity variables $\lambda^*$ obtained when solving (5), i.e., $I^*_{MPA} = \sum_{k=1}^{M} \lambda_k^* I_k$. Most likely, the MPA will not be an observed alternative, but rather a linear combination of those alternatives whose optimal lambdas are positive, $\lambda_k^* > 0$—in case an observed alternative $k$ is an MPA then $\lambda_k^* = 1$.

Note that the UMP searches for the MPA without imposing any restriction on the proportions that must be observed between the variables. In the case of the Leontief utility function, the preferences of the DM over the set of indicators through their weights determine the proportionality between indicators. If we redefine the utility function to incorporate the weights, i.e., $U^{LT}(I) = \min\{w_i I_i, i = 1, \ldots, N\}$, then the greater the weight $w_i$ the smaller the amount required of $I_i$ to reach a given utility level. This feature is common to the other two specifications, but in this case the weights define a fixed proportionality between the quantities of the indicators yielding equal utility, i.e., $\min\{w_i I_i\} = \min\{w_j I_j\} \Leftrightarrow w_i I_i = w_j I_j, \forall i \neq j, i, j = 1, \ldots, N$, or, alternatively, $w_i/w_j = I_j/I_i, \forall i \neq j, i, j = 1, \ldots, N$. For example, if $w_i = 0.75$ and $w_j = 0.25$, then $I_i = I_j/3$ and the amount of $I_i$ must be one third of $I_j$ (1:3).[7] Incorporating the fixed proportions defined by the DM requires including these bilateral proportions as constraints in program (5)—in DEA terminology the set of weight restrictions (see Ref. [55]). The number of restrictions corresponds to the combination of the $M$ variables taken two at a time without repetition, i.e., $C_2^M = M!/(2(M-2)!)$. However, once the required proportionality among weights and indicators is considered, all combinations simplify by substitution to the bilateral restrictions of one weight (e.g., that of the first variable) against the rest, so the number of restrictions is $M-1$. Under these restrictions it may be feasible that, when solving program (5), some of the restrictions $\sum_{k=1}^{M} \lambda_k I_{ki} \geq I_i, i = 1, \ldots, I$, are not saturated in the optimal solution (i.e., they are verified as inequalities), resulting in the existence of slacks.

**Proposition 2.** *If the utility function $U(I)$ is a continuous function, then the utility maximization problem has a solution.*

*Proof.* Mas-Colell et al. [13], p. 943). *Since the definition of the attainable set T in* (2) *ensures that it is a compact set* (*i.e., bounded and closed*), *then a continuous function always attains a maximum* [*Weierstrass Theorem*].

### 2.5. An example

We illustrate in Fig. 1 the UMP subject to the attainment set. We consider the previous three specifications of the utility function considering equal weights. Linear: $U^{LR}(I) = 0.5I_1 + 0.5I_2$; Cobb-Douglas: $U^{CD}(I) = I_1^{0.5} I_2^{0.5}$; and Leontief: $U^{LT}(I) = \min\{I_1, I_2\}$, subjeto to $I_1 = I_2$. In this latter case, the DM determines the maximum utility assuming that equal amounts of the indicators yield the same utility. That is, since $w_1 I_1 = w_2 I_2$, or, equivalently, $w_1/w_2 = I_2/I_1$, we have that $w_1/w_2 = 0.5/0.5 = 1$ and, therefore, $I_1 = I_2$.

Table 2 reports that the preferred alternative of the DM maximizing her utility under each utility function is B: i.e., $U^{LR}(I^*_B) = U^{CD^*}(I^*_B) = U^{LT^*}(I^*_B) = 0.8$ (shaded in grey in Table 2). For each utility specification the corresponding indifference contour associated with the MPA is represented in Fig. 1. The example is designed in such way that maximizing the utility under the above specifications and weights identifies the same benchmark (MPA) on the strongly effective frontier $F^S(T)$ and yields the same utility value. As shown in the empirical application this is not the case normally, and therefore we can make the following remarks:

*Different functions with equal weights identify alternative MPA and utility values.* As the UMP (5) depends on the attainable set and the formulation of the utility function, the benchmarks will differ across functions even under common weights. For instance, changing the weights to $w_1 = 0.75$ and $w_2 = 0.25$ results in:

- $U^{LR}(I^*_C) = 0.825$;

[7] Alternatively, if the weights were intended to reflect desirable proportions between the quantities of indicators, i.e., $w_i/w_j = I_i/I_j, \forall i \neq j, i, j = 1, \ldots, N$, the utility function would be inversely redefined as $U^{LT}(I) = \min\{I_i/w_i, i = 1, \ldots, N\}$. In the example this reverses the proportionality to 3:1.

**Table 2**
Benchmarking alternatives through utility functions, equal weights: $w_i = 1/N$.

| Alternatives | Observed Utility | | | Utility Effectiveness | | | | Utility Effectiveness Decompositions (strongly effective frontier, $F^S(T)$) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Linear | Cobb-Douglas | Leontief | Linear | Cobb-Douglas | Leontief | Tech. Effec. | Linear | | Cobb-Douglas | | Leontief | |
| $(I_1, I_2)$ | $U^{LR}(I)$ | $U^{CD}(I)$ | $U^{LT}(I)$ | $UE^{LR}$ | $UE^{CD}$ | $UE^{LT}$ | $1/\phi_k^*$ | Technical: $UTE_k^S$ | Allocative: $UAE_k^S$ | Technical: $UTE_k^S$ | Allocative: $UAE_k^S$ | Technical: $UTE_k^S$ | Allocative: $UAE_k^S$ |
| A = (0.3,1) | 0.650 | 0.548 | 0.300 | 0.813 | 0.685 | 0.375 | 1.000 | 1.000 | 0.813 | 1.000 | 0.685 | 1.000 | 0.375 |
| B = (0.8,0.8) | 0.800 | 0.800 | 0.800 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| C = (1,0.3) | 0.650 | 0.548 | 0.300 | 0.813 | 0.685 | 0.375 | 1.000 | 1.000 | 0.813 | 1.000 | 0.685 | 1.000 | 0.375 |
| D = (0.5,0.5) | 0.500 | 0.500 | 0.500 | 0.625 | 0.625 | 0.625 | 0.625 | 0.625 | 1.000 | 0.625 | 1.000 | 0.625 | 1.000 |
| E = (0.3,0.5) | 0.400 | 0.387 | 0.300 | 0.500 | 0.484 | 0.375 | 0.556 | 0.554 | 0.903 | 0.554 | 0.875 | 0.554 | 0.677 |
| F = (1,0.1) | 0.550 | 0.316 | 0.100 | 0.688 | 0.395 | 0.125 | 1.000 | 0.846 | 0.813 | 0.577 | 0.685 | 0.333 | 0.375 |
| $U(I_B^*)$ | 0.800 | 0.800 | 0.800 | – | – | – | | – | – | – | – | – | – |

Note: The decomposition of utility effectiveness considering the weakly and strongly effective frontiers, (11) and (14), are the same because all alternatives either belong to or are projected to the strongly efficient frontier (i.e., are slacks are zero). The only exception is alternative F, located on the weakly effective frontier, i.e., $\phi_F^* = 1$, and a positive slack in the second indicator, $s_{F2}^{+*} = 0.2$, the decomposition is equal to: $UE_F = UTE_F^W \times UAE_F^W = 1 \times UAE_F^W$, with $UE_F = UAE_F^W$ for each specification of the utility function.

- $U^{CD*}(I_{\overline{BC}}^*) = 0.8$, with the MPA being a linear combination of alternatives B and C. The intensity variables are $\lambda_B^* = 0.692$ and $\lambda_C^* = 0.308$ yielding $I_{MPA}^* = (0.862, 0.646)$ ;
- $U^{LT*}(I_{\overline{AB}}^*) = 0.329$, with the weights $w_1/w_2 = I_2/I_1 = 0.75/0.25 = 3$. Therefore, the DM determines that the utility of $I_1$ is three times larger than the utility of $I_2$ (e.g., $I_1 = 0.1$ yields the same utility as $I_2 = 0.3$). Solving model (5) with the constraint corresponding to this fixed proportion yields $I_{MPA}^* = (0.329, 0.988)$, which is the linear combination of alternatives A and B with $\lambda_B^* = 0.941$ and $\lambda_C^* = 0.059$.

*The MPA may belong to the weakly efficient frontier* $F^W(T)$. Let us consider the Leontief function with $w_1 = 0.8$ and $w_2 = 0.2$. Solving the UMP with the corresponding fixed proportion constraint yields $I_{MPA}^* = (0.25, 1)$, which belongs to the weakly efficient frontier $F^W(T)$. In the optimal solution $\lambda_A^* = 1$, and therefore the restriction corresponding to the first variable is not saturated, i.e., $\lambda_A^* I_{A1} > I^*$ $(0.3 > 0.25)$. Hence, the difference between the two is the slack existing in that dimension: $s_1^{+*} = 0.3 - 0.25 = 0.05$.

*Multiple optima.* Given that the attainability set is a convex polyhedral, conformed by intersecting hyperplanes, it is feasible that in the case of the linear utility function—and in combination with a particular set of weights—there may exist multiple solutions to the UMP. For instance, removing observation C would make segment AB the strongly effective frontier of the new attainability set. Under the linear function and equal weights, $w_1 = 0.5$ and $w_2 = 0.5$, all alternatives in this segment would maximize the utility.[8]

### 2.6. Benchmarking performance: utility effectiveness

Once the MPA has been determined by the DM, it is possible to benchmark the performance of the alternatives by defining the ratio of observed utility to maximum utility, thereby measuring the *utility effectiveness* of the alternatives. That is, for alternative $k$,

$$UE_k = \frac{U(I_k)}{U(I_{MPA}^*)} \leq 1, k = 1, \ldots, M. \quad (6)$$

An alternative is effective if it maximizes program (5), i.e., it is a MPA and $UE_k = 1$. Otherwise, if $UE_k < 1$ it is deemed ineffective, and $UE_k$ measures its relative utility effectiveness. For instance, $UE_k = 0.5$, implies that the alternative yields half the utility of the MPA. In this case, the *k-th* observation could double its performance in the eyes of the DM if it changed the value of its indicators to match those of the MPA. Therefore, the greater the value, the larger the utility effectiveness and vice versa. Table 2 reports the relative effectiveness of the alternatives with respect to B for each specification of the CES utility function. Fig. 1 illustrates utility effectiveness considering the linear specification: $U^{LR}(I) = 0.5I_1 + 0.5I_2$. For example, the maximum utility corresponding to $U^{LR}(I_B^*) = 0.8$ is represented by the green line tangent to the attainable set. Then the utility of alternatives A and C, $U^{LR}(I_A) = U^{LR}(I_C) = 0.65$, is represented by the brown dashed line, while their utility effectiveness, $UE_A = UE_C = 0.813 = 0.65/0.8$ corresponds to the gap between the brown and green lines. Similarly, for alternative E, its utility effectiveness $UE_E = 0.5 = 0.4/0.8$ is represented by the gap between the red and green lines.

The disparity in the range of values for utility effectiveness depends on the choice of utility function. More importantly, the ranking of alternatives under each utility function differs beyond the MPA. For example, under the linear and Cobb-Douglas functions, the second most preferred alternatives are A and C, but under the Leontief function it is alternative D. After that, each utility function has its own ranking. The relevance of the rankings is paramount since DMs use them to categorize the alternatives. For example, the EIS uses a linear utility function with equal weights to define the so-called 'Summary Innovation Index' (SII). Based on the values of the SII, they classify countries as 'Innovation leaders', 'Strong innovators', 'Moderate innovators' and 'Emerging innovators' [2]. Changing the utility functions or the weights would result in alternative rankings, with countries classified differently. This shows the subjective nature of any evaluation process when ranking the alternatives according to their performance under a specific aggregating utility function and set of weights.

A practical way to choose among different utility specifications is to treat them as alternative assumptions about substitutability across indicators, and select the one that best matches the decision context (or use them as a sensitivity check when that context is unclear). Linear utility implies perfect substitutability (strong compensation: high performance in one dimension can fully offset low performance in another); Leontief implies no substitutability (non-compensatory: weak performance in any dimension strongly limits overall performance); and Cobb–Douglas/CES lies in between (partial, smooth substitutability with

[8] Imposing the property of strict quasi-concavity of the utility function $U(I)$, as with the Cobb-Douglas and Leontief specifications, ensures that the preferences of the DM are strictly convex, preventing the existence of multiple optima in combination with a polyhedral attainability set ([13]; Ch. 3).

diminishing returns). In many policy benchmarking applications, including the empirical section of this study on innovation systems, the true degree of compensation is unknown (although the EC favors the perfect substitutability imposed by the simplest utility function corresponding to the arithmetic mean), so reporting results for the three cases is not a drawback but a robustness/sensitivity exercise that brackets realistic preference structures—showing how rankings and effectiveness change from fully compensatory to fully non-compensatory aggregation.

## 3. Decomposing utility effectiveness into technical (quantity) and allocative (weight) factors

It is possible to shed light on the sources of utility ineffectiveness with respect to the most preferred benchmark by considering the performance of an alternative within the attainable set. The rationale is that, when evaluating how effective alternatives are, the DM determines first if the alternative under evaluation belongs to the frontier of the attainable set; in particular, the strongly efficient frontier defined by observations that are Pareto effective. The reason is clear, a dominated alternative cannot maximize utility regardless of the specific function and weights chosen by the DM. Therefore, preference weights are irrelevant at this stage of the analysis, which is purely technical and based on the quantity of the variables. Any DM interested in improving the performance of an alternative expects it to increase the quantity of the indicators to reach the effective frontier to increase its utility. Arguably, the easiest way to increase effectiveness is to expand the value of the indicators in their existing proportions, implying that the structure of the system being evaluated remains unchanged. For instance, considering that the current structure of an innovation system is the result of past investment decisions, the EC can simultaneously recommend increasing the population with tertiary education and R&D expenditure in the public sector by 20% without changing the fixed proportions among these indicators.

One can resort to DEA to determine the technical effectiveness of an alternative through the so-called *quantity (technical) effectiveness* score measuring the radial distance from an alternative to the weakly frontier $F^W(T)$.[9] Linear program (7) below operationalizes the above proposal by calculating the amount (factor) in which the vector of observed indicators can be radially expanded until a projection on $F^W(T)$ is achieved. However, since the benchmark alternative on the frontier may be dominated in individual dimensions, a second program is solved to determine if any slacks exist. These individual slacks, denoted by $s_i^+$, $i=1, \ldots, N$, are calculated through program (8). This ensures that any inefficient alternative is projected to a benchmark that is Pareto-effective by belonging to the strongly effective frontier, $F^S(T)$. Therefore, denoting by $o$ the alternative under evaluation we solve the following linear programs $M$ times, one for each alternative[10]

First stage: Radial output technical effectiveness

$$\max_{\phi,\widehat{\lambda}} \ \phi_o$$

$$\begin{aligned} &\text{s.t.} \\ &\sum_{k=1}^{M} \widehat{\lambda}_k I_{ki} \geq \phi_o I_{oi}, i=1,\ldots,N, \\ &\sum_{k=1}^{M} \widehat{\lambda}_k = 1, k=1,\ldots,K, \\ &\widehat{\lambda} \geq 0. \end{aligned} \tag{7}$$

Second stage: Additive slacks in effectiveness

$$\max_{s^+,\widetilde{\lambda}} \sum_{i=1}^{N} s_{oi}^+$$

$$\begin{aligned} &\text{s.t.} \\ &\sum_{k=1}^{M} \widetilde{\lambda}_k I_{ki} - s_{oi}^+ = \phi_o^* I_{oi}, i=1,\ldots,N, \\ &\sum_{k=1}^{M} \widetilde{\lambda}_k = 1, \ k=1,\ldots,K, \\ &\widetilde{\lambda} \geq 0, \quad s_o^+ \geq 0. \end{aligned} \tag{8}$$

The optimal solutions to the above programs (7) and (8) are denoted by $\phi_o^* \geq 1$ and $s_o^{+*} \geq 0$, respectively. An alternative is technically effective if $\phi_o^* = 1$ and $s_o^{+*} = 0$. Otherwise, it does not belong to the strongly effective frontier—although it belongs to the weakly efficient frontier if $\phi_o^* = 1$ and $s_o^{+*} > 0$. From program (7) one obtains the projection on the weakly effective frontier, denoted by $\widehat{I}_o^W$, as the linear combination of the alternatives that serve as benchmark for the alternative under evaluation, i.e., $\widehat{I}_o^W = \phi^* I_o = \sum_{k=1}^{M} \widehat{\lambda}_k^* I_{ki}$. Building on this solution, program (8) searches for the individual slacks $s_{oi}^+$, $i=1, \ldots, I$, once the alternative $I_o$ under evaluation has been projected to the weakly efficient frontier—as shown by the inclusion of $\phi_o^* I_o$ in the right hand side of the $N$ restrictions. Then, using the solutions to programs (7) and (8), it is possible to determine the reference alternative on the strongly effective frontier of the attainable set, denoted by $\widehat{I}_o^S$, by calculating first the radial projection of alternative $I_o$ to the weakly efficient frontier, $\widehat{I}_o^W = \phi^* I_o$, and subsequently adding the vector of individual slacks. Thereby, the technically efficient benchmark corresponds to $\widehat{I}_o^S = \widehat{I}_o^W + s_o^{+*} = \phi_o^* I_o + s_o^{+*}$. Given that $\phi_o^* \geq 1$ and $s_o^{+*} \geq 0$, the following inequalities are verified for any alternative $k$: $\widehat{I}_k^S \geq \widehat{I}_k^W \geq I_k$.

Once the effectiveness of an alternative based on quantities has been calculated, it is possible to decompose utility ineffectiveness (6) considering either the weakly efficient frontier $\phi^* \geq 1$, or the strongly effective frontier, $s^{+*} \geq 0$, as we show next.

### 3.1. Decomposing utility effectiveness against the weakly effective frontier

Taking the weakly efficient frontier as benchmark represents the standard practice in DEA-BoD effectiveness models. This implies that the DM ignores the existence of slacks associated with the attainable set $T$ and is satisfied with benchmarking alternatives against peers that may be Pareto dominated. In this case the benchmark alternative corresponds to $\widehat{I}_o = \phi^* I_o$, and recalling that the CES utility function (1) is homogeneous of degree one, we observe that $U(\widehat{I}_k^W) = U(\phi_k^* I_k) = \phi_k^* U(I_k)$. Mirroring the ratio form of expression (6) we can define the following measure of *utility technical (quantity) effectiveness*:

$$UTE_k^W = \frac{U(I_k)}{U(\widehat{I}_k^W)} = \frac{U(I_k)}{\phi_k^* U(I_k)} = \frac{1}{\phi_k^*} \leq 1, k=1,\ldots,M. \tag{9}$$

Hence, if $UTE_k^W < 1$, the evaluated alternative is Pareto-dominated in quantities and does not belong to the weakly effective frontier, resulting in a loss of utility that can be measured in proportional terms through the (inverse) of the radial measure $\phi_k^*$. For instance, $\phi_k^* = 2$ indicates that the alternative yields half the utility of its projection on the frontier $F^W(T)$, which could be attained by doubling the amount of all indicators

[9] The radial model searching for feasible increases of all indicators in the same proportion corresponds to the output oriented canonical model in DEA [51]. Other popular effectiveness measures used in in BoD models are the additive model [15] and the directional distance function [17].

[10] With model (7) we follow Lovell and Pastor [52] and calculate an output-oriented variable returns to scale model without inputs. The BoD effectiveness model put forward by Cherchye et al. [14] is formulated as an input-oriented constant returns to scale DEA model in the multiplier form with a common single constant input with a value at one. Van Puyenbroeck [56] shows that there is a direct, reciprocal relation between both formulations.

simultaneously. On the contrary, if $UTE_k^W = 1$, the alternative is technically effective and cannot increase the utility of the DM by radially expanding the quantity of indicators.

One sees that DEA-BoD effectiveness ($UTE_k^W$) measures whether an alternative is technically effective, that is, whether it lies on the efficient frontier when it is allowed to choose the most favorable multipliers ($\widehat{\lambda}_k^*$). Utility effectiveness ($UE_k$) extends this concept by evaluating whether the technically effective benchmark also maximizes the utility of the DM, given a choice of utility function with a specific substitutability and weight structure. In this sense, DEA-BoD effectiveness captures technical feasibility—as it depends on quantities only, while utility effectiveness assesses allocative optimality relative to the DM's preferences; an alternative may be DEA-BoD effective but not utility-effective if the multipliers that ensure technical effectiveness differ from those that maximize utility.

Therefore, even if a (projected) alternative is technically effective by belonging to the weakly effective frontier, it may fall short of maximizing the utility of the DM. The reason is that the utility function of the DM does not identify alternative $k$ as MPA, implying that the allocation (mix or balance) of indicators, $\widehat{I}_k = (\widehat{I}_{1k}, \cdots, \widehat{I}_{Nk})$, is suboptimal given the structure of the preferences. From Proposition 2, positing the existence of a solution to the utility maximization problem (5), the following inequality is verified $U(I^*_{MPA}) \geq U(\widehat{I}_k)$. This implies that we can define the following measure of relative *utility allocative (weight) effectiveness* capturing the utility gap between an effective alternative and the most preferred one:

$$UAE_k^W = \frac{U(\widehat{I}_k^W)}{U(I^*_{MPA})} = \frac{\phi_k^*\ U(I_k)}{U(I^*_{MPA})} = \phi_k^*\ UE_k \leq 1, k = 1, ..., M, \tag{10}$$

where the last equality is obtained substituting the definition of utility effectiveness (6). As before, $UAE_k^W < 1$ indicates that the effective alternative on the weakly efficient frontier does not maximize the utility of the DM, while $UAE_k^W = 1$ signals that it is the MPA.[11] This can be easily shown by considering the optimal solutions to problems (7) and (5). On the one hand, the solution to program (7) yields the optimal multipliers $\widehat{\lambda}_k^*$ that identify the benchmark composite indicator on the frontier: $\widehat{I}_k^W = \sum_{k=1}^M \widehat{\lambda}_k^* I_{ki}$, and whose utility is calculated as: $U(\widehat{I}_k^W) = U\left(\sum_{k=1}^M \widehat{\lambda}_k^* I_{ki}\right)$. On the other hand, the alternative that maximizes the utility is the solution to program (5), yielding the optimal multipliers $\lambda_k^*$ that identify the most preferred linear combination of frontier indicators: $I^*_{MPA} = \sum_{k=1}^M \lambda_k^* I_{ki}$, and whose utility is $U(I^*_{MPA}) = U\left(\sum_{k=1}^M \lambda_k^* I_{ki}\right)$. Since the utility function (1) of the DM (defined by the aggregating functional form $U$, $\rho$ and weights, $w_i$), is the same for both indicators, $\widehat{I}_k^W$ and $I^*_{MPA}$, we see immediately that the only difference between the utilities of both indicators is the set of optimal multipliers, $\widehat{\lambda}_k^*$ and $\lambda_k^*$. Then, when $\widehat{\lambda}_k^* = \lambda_k^*$ the projection of alternative $I_k$, $\widehat{I}_k^W$, happens to maximize the utility of the DM, i.e., $\widehat{I}_k^W = I^*_{MPA}$, and therefore $U(\widehat{I}_k^W) = U(I^*_{MPA})$ resulting in $UAE_k^W = 1$. Alternatively, if the optimal multipliers obtained when solving (7) are different from those obtained when solving (5), i.e., $\widehat{\lambda}_k^* \neq \lambda_k^*$, then $\widehat{I}_k^W \neq I^*_{MPA}$ with $U(\widehat{I}_k^W) < U(I^*_{MPA})$ and, therefore, $UAE_k^W < 1$.

The maximization of the utility allows us to bring the preferences of the DM into the analysis rather than restricting it to simply determine if an alternative can increase its technical effectiveness. Note that from a DEA-BoD perspective, both composite indicators, $\widehat{I}_k^W$ and $I^*_{MPA}$, would have the same unitary solution when solving program (7), $\phi_o^* = 1$, meaning that they are effective from a technical perspective, and it is not possible to discriminate (or rank) between the two alternatives. Hence, utility allocative effectiveness $UAE_k^W$ allows establishing if an alternative maximizes the utility beyond its technical status (i.e., beyond the DEA-BoD model). This depends on the utility function of choice describing the preferences of the DM with respect to the degree of substitutability between indicators, $\rho$ (as presented in Table 1), and the weights $w_i$.

In case the alternative is ineffective from an allocative perspective, it must change its indicators to match the quantities of the MPA. Rearranging expression (10), we obtain the following decomposition of utility effectiveness (6):

$$UE_k = \frac{U(I_k)}{U(I^*_{MPA})} = \frac{1}{\phi_k^*} \times \frac{U(\widehat{I}_k^W)}{U(I^*_{MPA})} = UTE_k^W \times UAE_k^W \leq 1, k = 1, ..., M. \tag{11}$$

Therefore, for each alternative $k$ we can learn about the causes of utility ineffectiveness, being either technical (i.e., related to insufficiently valued indicators with respect to the weakly effective frontier of the attainable set, $UTE_k^W$) and/or allocative (i.e., related to wrong balance of indicators given the weight structure of the decision maker, $UAE_k^W$).

### 3.2. Decomposing utility effectiveness against the strongly effective frontier

Taking into consideration the existence of individual slacks that allow for non-proportional increases in the indicators leads to the decomposition of utility effectiveness against the strongly effective frontier. This implies that both radial and non-radial increments in the indicators are taken into account (i.e, $\widehat{I}_k^S = \phi_k^* I_k + s_k^{+*}$). In this case the homogeneity property of the utility function cannot be invocated, and we define *utility technical (quantity) effectiveness* against $F^S(T)$ as follows:

$$UTE_k^S = \frac{U(I_k)}{U(\widehat{I}_k^S)} = \frac{U(I_k)}{U(\phi_o^* I_o + s_o^{+*})} \leq 1, k = 1, ..., M, \tag{12}$$

where the numerical values are interpreted in the same way: $UTE_k^S < 1$ indicates that the evaluated alternative does not belong to the strongly effective frontier, resulting in a loss of utility. On the contrary, if $UTE_k^S = 1$, the alternative is technically effective and cannot increase the utility of the DM by expanding (both radially and non-radially) the quantity of indicators.

However, as before, a technically effective alternative may fall short of maximizing the DM's utility given the structure of weights of her utility function; i.e, $U(I^*_{MPA}) \geq U(\widehat{I}_k^S)$. In this case we define the relative *utility allocative (weight) effectiveness* as follows:

$$UAE_k^S = \frac{U(\widehat{I}_k^S)}{U(I^*_{MPA})} = \frac{U(\phi_k^* I_k + s_k^{+*})}{U(I^*_{MPA})} \leq 1, k = 1, ..., M, \tag{13}$$

with the same interpretation that its weakly efficient counterpart (10). Combining the last two expressions we obtain the corresponding decomposition of utility effectiveness (6):

$$UE_k = \frac{U(I_k)}{U(I^*_{MPA})} = \frac{U(I_k)}{U(\widehat{I}_k^S)} \times \frac{U(\widehat{I}_k^S)}{U(I^*_{MPA})} = UTE_k^S \times UAE_k^S \leq 1, k = 1, ..., M. \tag{14}$$

Again, technical ineffectiveness refers to insufficiently valued (in quantity) indicators with respect to the strongly effective frontier of the attainable set, $UTE_k^S$, while allocative ineffectiveness, $UAE_k^S$ is related to a wrong balance of indicators given the weight structure of the DM's utility function.

Comparing the decompositions based on the strongly and weakly effective frontiers, we observe that the following inequalities hold:

[11] Note that, as shown in the example, under a Leontief utility specification, an alternative may be the MPA even if slacks exist in individual dimensions.

$UTE_k^S \leq UTE_k^W$ and, therefore, $UAE_k^S \geq UAE_k^W$, with the two decompositions being equal when the alternative under evaluation is projected radially through (7) onto the strongly effective frontier and, therefore, the optimal slacks are null, $s_o^{+*} \geq 0$.

### 3.3. An example (cont.)

Table 2 presents the decomposition of utility effectiveness $UE_k$ with respect to the MPA B under each utility specification and considering the strongly effective frontier $F^S(T)$ as reference. All alternatives, except F, either belong or are radially projected through program (7) to this frontier (e.g., alternative E is projected to the linear combination E′ of alternatives A and B in Fig. 1, implying that their associated optimal slacks in program (8) are zero). Consequently, for these alternatives, A thru E, the decompositions associated with the weakly and strongly effective frontiers are equal. Moreover, since there are no slacks, the three specifications yield the same utility technical effectiveness $UTE_F^S$, which is equal to the (inverse) of their effectiveness score, $1/\ \phi_k^*$ (reported in the 8th column under 'Tech. Effec.'), given the homogeneity property of the CES function. For this reason, it is the associated allocative factor, $UAE_F^S$, the term capturing the difference between the different utility effectiveness corresponding to the linear, Cobb-Douglas and the Leontief specifications.

However, in applications with real data the existence of positive slacks is pervasive. This implies that the three decompositions have different values when considering the strongly effective frontier as reference. This can be illustrated with alternative F, which belongs to the weakly effective frontier, with an effectiveness score equal to $\phi_F^* = 1$, and a positive slack in the second indicator, $s_{F2}^{+*} = 0.2$. As for any alternative that belongs or is projected to the weakly effective frontier, the three technical effectiveness factors are different. For the linear utility function we obtain: $UTE_F^S|_{LR} = 0.846$, while for the Cobb-Douglas and Leontief functions these are $UTE_F^S|_{CD} = 0.577$ and $UTE_F^S|_{LT} = 0.333$. The residual nature of the allocative terms makes them dependent on $UE_k$ and $UTE_F^S$, and therefore their values adjust accordingly to complete the decompositions. In the empirical application evaluating the utility effectiveness of national innovation systems we show the practical relevance of measuring utility effectiveness against Pareto effective benchmarks, i.e., the strongly effective frontier, so the sources of utility ineffectiveness are correctly identified by DMs.

Finally, we remark that the difference between the strongly and weakly effective frontiers emerges in the context of standard applications of the DEA model using programs (7) and (8). As surveyed in the introduction, the DEA and BoD literature offers alternatives that might result in less frequent slacks. For instance, it is possible to impose either most preferred solutions —following value efficiency analysis—or, equivalently, a set of restrictions on the weights of the program dual to (7), which can prevent the existence of slacks by removing hyperplanes with zero weights (and positive slacks). For example, considering the example in Fig. 1, this would result in a reconfiguration of the strongly efficient frontier whose two supporting hyperplanes would be those defined by alternatives A and B, $\overline{AB}$, and B and C, $\overline{BC}$, and their extensions to the axes. This effectively removes the horizontal and vertical dashed extensions corresponding to the weakly efficient frontier. Nevertheless, the adoption of these restrictions implies further normative judgements made by the DM, along with the preferred utility function.

## 4. Utility effectiveness of national innovation systems

### 4.1. Data and sources: the European Innovation Scoreboard

The approach presented in this study is used to evaluate the effectiveness of national innovation systems considering the 2024 edition of the EIS.[12] The EIS provides a series of 32 indicators that aim at evaluating the innovation systems of the 27 EU member countries—online Appendix lists all 32 indicators. To make our results comparable to those obtained by the European Commission [2] we adopt their weighting scheme, which assigns equal values to all indicators. Previous studies, such as Edquist et al. [57], Barbero et al. [43] and Zofío et al. [58], have relied on EIS data to analyze the performance of innovation systems from an efficiency perspective by grouping the EIS indicators into input and output categories. In this study all indicators are categorized as outputs to illustrate the new methodology combining CIs and DEA-BoD. The EIS encompasses 12 dimensions (sub-pillars) ranging from 'Human resources' to 'Environmental sustainability', which are further aggregated into four main types of activities (pillars): 'Framework conditions', 'Investments', 'Innovation activities', and 'Impacts'. Due to the high number of indicators, classical DEA methods run into the so-called curse of dimensionality, which results in almost all countries being classified as efficient. For this reason, researchers tend to either group indicators in higher categories (as the aforementioned 12 dimensions or four types), select a representative subset of indicators before running the DEA models, or apply multivariate or regression methods to reduce their dimensionality.

### 4.2. 'Ensemble-DEA' methods and the DEA-BoD model

However, the fact that the standard DEA-BoD method proves insufficient to measure effectiveness in studies involving a high number of variables relative to the number of observations has been solved through the so-called 'Ensemble-DEA'. Recently introduced by Aparicio et al. [38], this methodology adopts ensemble-based techniques consisting of double-randomization of the data, applied to observations and variables. By retaining all observations and variables in the original sample, the method maintains its integrity while retaining the interpretability of the efficiency scores. More specifically, Ensemble-DEA generates a sufficiently large number of different subsamples of the original data matrix by randomly replacing alternatives (rows) and indicators (columns) in such a way that a consistent balance between them is preserved, thereby avoiding the curse of dimensionality. The method is a crossover between model ensembles and randomization with respect to observations and indicators, effectively using strategies commonly employed in machine learning and data science and providing a more robust framework for efficiency analysis in general. Finally, as shown in our empirical application, the Ensemble-DEA method can offer detailed benchmarking information, including the identification of peers for each observation evaluated, providing DMs with important insights to drive performance improvements. The method consists of the following steps:

1) *Step 1. Determination of the maximum number of indicators.* Following standard practice, the size of a representative subsample balancing observations and indicators is determined through the heuristic rule proposed by Cooper et al. [59]. This popular rule establishes that the maximum number of indicators, $N$, relative to a number of alternatives, $M$, must satisfy $M \geq 3\cdot(N+1) \Leftrightarrow N \leq \frac{M}{3} - 1$. Consequently, since the number of countries is given in our application, we define the maximum number of indicators $N^{\#}$ to be used for each subsample as the nearest integer to $\frac{M}{3} - 1$. In our case, since $M = 27$, then $N^{\#} = 8$.
2) *Step 2. Generation of bootstrap samples and application of random alternative and indicator selection.* Once the size of the subsample has been determined (27 countries and 8 indicators in our case), the next step is the generation of $P$ bootstrap samples, denoted as $\aleph_q$, $q = 1, \ldots, P$. The process entails generating several subsets of the original dataset through country and indicator resampling (note that country

[12] European Commission [2] offers a complete discussion of the statistical sources and variable specifications.

$k$ can be repeated in sample $\aleph_q$). This method guarantees that all dimensions of the dataset are included in the analysis if $P$ is sufficiently large, thereby preventing any single country $k$ or any single indicator $i$ from dominating the results.

3) *Step 3. Calculation of the Utility Effectiveness measure $UE_k(\aleph_q)$ for the bootstrap samples.* For each sample $\aleph_q$ we calculate country's $k$ utility effectiveness $UE_k(\aleph_q)$ in accordance with (6) and using the different linear, Cobb-Douglas and Leontief specifications. This requires computing countries' utilities, $U(I_k(\aleph_q))$, as well as the maximum utility archived by the MPA in the sample, $U(I^*_{MPA}(\aleph_q))$, obtained by solving problem (5). As for the weights, we maintain the same neutral scheme adopted by the EC, i.e. $w_i = 1/N^{\#} = 0.125$, $i = 1, \ldots, N$
4) *Step 4. Calculation of the weak and strong Utility Technical Effectiveness measures, $UTE^W_k(\aleph_q)$ and $UTE^S_k(\aleph_q)$, for the bootstrap samples.* Measuring utility technical effectiveness against the weakly efficient frontier (4) requires to calculate country's $k$ effectiveness score solving program (7) for each sample $\aleph_q$, i.e., $UTE^W_k(\aleph_q) = 1/\phi^*_k$. Additionally, determining utility technical effectiveness against the strongly efficient frontier $UTE^S_k(\aleph_q)$ requires solving program (8) for any existing slacks, $s^{+*}_k$. Subsequently, the utility of the projection is calculated, $\widehat{I}^S_k) = U(\phi^*_k I_k + s^{+*}_k)$ , which in turns allows determining $UTE^S_k(\aleph_q)$ in accordance with (12).
5) *Step 5. Decomposition of the Utility Effectiveness measure $UE_k(\aleph_q)$ against the weakly and strongly efficient frontiers.* Utility effectiveness can be decomposed into technical and allocative effectiveness against the weakly and strongly efficient frontiers following expressions (11) and (14), respectively. This requires recovering, for each sample $\aleph_q$, the allocative terms $UAE^W_k(\aleph_q)$ and $UAE^S_k(\aleph_q)$.
6) *Step 6. Aggregation of the effectiveness measures for each alternative.* To summarize the results across the $P$ solved bootstrap samples we aggregate individual effectiveness using the so-called denominator rule. Balk and Zofio [60] study existing rules that ensure a consistent aggregation of measurement scores whereby the aggregate retains the same interpretation and properties of the individual measures. Different rules can be defined using either the (weighted) arithmetic, geometric or harmonic means. Here we choose the popular arithmetic mean whose weights correspond to the shares of the maximum utility value achieved in each sample $\aleph_q$ over the sum of all maximum utility values across all $P$ samples (for each sample $\aleph_q$ accounting for the number of times a country $k$ is present, $k \in \aleph_q$); i.e.,

$$\overline{UE}_k = \frac{\sum_{\substack{q=1,\ldots,P \\ k\in\aleph_q}} U(I_k(\aleph_q))}{\sum_{\substack{q=1,\ldots,P \\ k\in\aleph_q}} U(I^*_{MPA}(\aleph_q))} = \sum_{\substack{q=1,\ldots,P \\ k\in\aleph_q}} \frac{U(I^*_{MPA}(\aleph_q))}{\sum_{\substack{l=1,\ldots,P \\ k\in\aleph_q}} U(I^*_{MPA}(\aleph_l))} \left(\frac{U(I_k(\aleph_q))}{U(I^*_{MPA}(\aleph_q))}\right) \tag{15}$$

Then, aggregate utility effectiveness can be decomposed in a consistent manner into aggregate utility technical effectiveness and aggregate utility allocative effectiveness as follows:

$$\overline{UE}_k = \underbrace{\frac{\sum_{\substack{q=1,\ldots,P \\ k\in\aleph_q}} U(I_k(\aleph_q))}{\sum_{\substack{q=1,\ldots,P \\ k\in\aleph_q}} U(\widehat{I}^F_k(\aleph_q))}}_{\overline{UTE}^F_k} \times \underbrace{\frac{\sum_{\substack{q=1,\ldots,P \\ k\in\aleph_q}} U(\widehat{I}^F_k(\aleph_q))}{\sum_{\substack{q=1,\ldots,P \\ k\in\aleph_q}} U(I^*_{MPA}(\aleph_q))}}_{\overline{UAE}^F_k} =$$

$$= \underbrace{\sum_{\substack{q=1,\ldots,P \\ k\in\aleph_q}} \frac{U(\widehat{I}^F_k(\aleph_q))}{\sum_{\substack{l=1,\ldots,P \\ k\in\aleph_q}} U(\widehat{I}^F_k(\aleph_q))} \left(\frac{U(I_k(\aleph_q))}{U(\widehat{I}^F_k(\aleph_q))}\right)}_{\overline{UTE}^F_k} \tag{16}$$

$$\times \underbrace{\sum_{\substack{q=1,\ldots,P \\ k\in\aleph_q}} \frac{U(I^*_{MPA}(\aleph_q))}{\sum_{\substack{l=1,\ldots,P \\ k\in\aleph_q}} U(I^*_{MPA}(\aleph_q))} \left(\frac{U(\widehat{I}^F_k(\aleph_q))}{U(I^*_{MPA}(\aleph_q))}\right)}_{\overline{UAE}^F_k},$$

where $F$=$W$,$S$, depending on whether the alternative under evaluation is projected to the weakly or strongly efficient frontier: $\widehat{I}^F_k(\aleph_q)$.

### 4.3. Results

Following this step-by-step methodology we apply 'Ensemble-DEA' to calculate utility technical effectiveness models (7) and (8). We run a total of $P$=50,000 subsamples because this number is large enough for the results to converge to a steady value. In each subsample we consider eight indicators (see Step 1 above and the adoption of Cooper et al.'s [59] 'rule of thumb') and 27 countries. Table 3 ranks EU countries by the average value of their observed linear utilities $\overline{U}^{LR}(I)$. The ranking mirrors that offered by the EIS since the linear utility $U^{LR}(I)$ is equal to the SII reported by the EIS. Denmark and Romania are the best and worst performing countries. Denmark constitutes the MPA for all utility specifications, and triples in value the utility of Romania, 0.750 versus 0.241. The utility values and rankings show a high degree of correlation between the average and geometric means despite the different substitutability assumptions: $\rho(\overline{U}^{LR}(I), \overline{U}^{CD}(I)) = 0.992$, but differences emerge when fixed proportions are assumed $\rho(\overline{U}^{LR}(I), \overline{U}^{LT}(I)) = 0.886$, $\rho(\overline{U}^{CD}(I), \overline{U}^{LT}(I)) = 0.933$. We observe that Proposition 1 is verified for each individual country and on average for all countries, with the fixed proportion values being substantially smaller than the linear and geometric aggregators. The assumption of fixed proportions can lead to substantial changes in the ranking of performance. For instance, Finland ranks 3rd according to $\overline{U}^{LR}(I_{Finland}) = 0.706$, but 19th according to $\overline{U}^{LT}(I_{Finland}) = 0.075$, which represents a relevant drop and shows that the Finnish innovation system is quite unbalanced with respect to the assessment scheme represented by the equal weights.

#### 4.3.1. Utility effectiveness and its decomposition

Utility effectiveness is calculated using definition (6) dividing the utility of each country $k$ by maximum utility. Showing consistency with the observed utility values in all three specifications, Denmark and Romania are the countries with the highest and lowest utility

**Table 3**
Benchmarking European innovation systems, equal weights: $w_i = 1/8$. Average aggregates of 50,000 samples.

| Country | Observed Utility | | | Utility Effectiveness | | | | Utility Effectiveness Decompositions (Strongly Effective Frontier, $I^S(T)$) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Linear | Cobb-Douglas | Leontief | Linear | Cobb-Douglas | Leontief | Tech. Effec. | Linear | | Cobb-Douglas | | Leontief | |
| $(I_1,\ldots,I_8)$ | $\overline{U}_k^{LR}(I)$ | $\overline{U}_k^{CD}(I)$ | $\overline{U}_k^{LT}(I)$ | $\overline{UE}^{LR}$ | $\overline{UE}^{CD}$ | $\overline{UE}^{LT}$ | $1/\phi_k^*$ | Technical: $\overline{UTE}_k^S\|_{LR}$ | Allocative: $\overline{UAE}_k^S\|_{LR}$ | Technical: $\overline{UTE}_k^S\|_{CD}$ | Allocative: $\overline{UAE}_k^S\|_{CD}$ | Technical: $\overline{UTE}_k^S\|_{LT}$ | Allocative: $\overline{UAE}_k^S\|_{LT}$ |
| Denmark | 0.749 | 0.727 | 0.495 | 0.954 | 0.950 | 0.749 | 0.999 | 0.998 | 0.956 | 0.998 | 0.953 | 0.997 | 0.751 |
| Sweden | 0.734 | 0.696 | 0.379 | 0.940 | 0.915 | 0.583 | 0.999 | 0.997 | 0.943 | 0.996 | 0.919 | 0.990 | 0.588 |
| Finland | 0.705 | 0.653 | 0.335 | 0.906 | 0.861 | 0.517 | 0.997 | 0.986 | 0.919 | 0.980 | 0.879 | 0.951 | 0.543 |
| Netherlands | 0.695 | 0.666 | 0.399 | 0.893 | 0.880 | 0.616 | 0.989 | 0.970 | 0.922 | 0.965 | 0.911 | 0.932 | 0.661 |
| Belgium | 0.683 | 0.657 | 0.408 | 0.878 | 0.867 | 0.629 | 0.988 | 0.967 | 0.908 | 0.964 | 0.899 | 0.945 | 0.666 |
| Austria | 0.643 | 0.627 | 0.455 | 0.828 | 0.830 | 0.697 | 0.981 | 0.946 | 0.874 | 0.944 | 0.879 | 0.935 | 0.746 |
| Ireland | 0.627 | 0.562 | 0.240 | 0.807 | 0.743 | 0.370 | 0.986 | 0.962 | 0.839 | 0.951 | 0.782 | 0.881 | 0.420 |
| Luxembourg | 0.620 | 0.553 | 0.239 | 0.797 | 0.731 | 0.369 | 0.987 | 0.956 | 0.834 | 0.945 | 0.774 | 0.882 | 0.419 |
| Germany | 0.617 | 0.582 | 0.324 | 0.794 | 0.771 | 0.500 | 0.983 | 0.942 | 0.843 | 0.934 | 0.825 | 0.880 | 0.568 |
| Cyprus | 0.587 | 0.510 | 0.200 | 0.755 | 0.675 | 0.309 | 0.987 | 0.938 | 0.805 | 0.920 | 0.734 | 0.824 | 0.375 |
| Estonia | 0.579 | 0.542 | 0.295 | 0.746 | 0.718 | 0.455 | 0.948 | 0.862 | 0.865 | 0.846 | 0.848 | 0.753 | 0.605 |
| France | 0.575 | 0.541 | 0.296 | 0.741 | 0.716 | 0.458 | 0.939 | 0.852 | 0.870 | 0.832 | 0.861 | 0.701 | 0.653 |
| Slovenia | 0.503 | 0.478 | 0.282 | 0.648 | 0.633 | 0.438 | 0.860 | 0.750 | 0.864 | 0.732 | 0.865 | 0.622 | 0.704 |
| Spain | 0.497 | 0.466 | 0.281 | 0.640 | 0.617 | 0.434 | 0.878 | 0.756 | 0.846 | 0.735 | 0.839 | 0.642 | 0.676 |
| Czechia | 0.496 | 0.464 | 0.258 | 0.639 | 0.615 | 0.400 | 0.867 | 0.766 | 0.834 | 0.744 | 0.826 | 0.621 | 0.643 |
| Italy | 0.495 | 0.469 | 0.288 | 0.638 | 0.622 | 0.446 | 0.899 | 0.772 | 0.827 | 0.758 | 0.819 | 0.687 | 0.649 |
| Malta | 0.486 | 0.389 | 0.133 | 0.627 | 0.515 | 0.205 | 0.938 | 0.830 | 0.755 | 0.767 | 0.672 | 0.544 | 0.377 |
| Lithuania | 0.462 | 0.411 | 0.199 | 0.595 | 0.544 | 0.308 | 0.881 | 0.754 | 0.789 | 0.721 | 0.755 | 0.606 | 0.509 |
| Portugal | 0.461 | 0.425 | 0.227 | 0.594 | 0.563 | 0.351 | 0.824 | 0.691 | 0.859 | 0.659 | 0.854 | 0.507 | 0.692 |
| Greece | 0.428 | 0.362 | 0.145 | 0.552 | 0.480 | 0.225 | 0.853 | 0.716 | 0.771 | 0.659 | 0.729 | 0.455 | 0.494 |
| Hungary | 0.390 | 0.354 | 0.179 | 0.502 | 0.469 | 0.278 | 0.725 | 0.613 | 0.819 | 0.580 | 0.808 | 0.436 | 0.637 |
| Croatia | 0.384 | 0.331 | 0.126 | 0.495 | 0.438 | 0.196 | 0.745 | 0.560 | 0.885 | 0.499 | 0.879 | 0.290 | 0.674 |
| Poland | 0.364 | 0.326 | 0.153 | 0.469 | 0.432 | 0.237 | 0.710 | 0.536 | 0.875 | 0.492 | 0.879 | 0.320 | 0.741 |
| Slovakia | 0.360 | 0.330 | 0.192 | 0.463 | 0.437 | 0.297 | 0.693 | 0.582 | 0.797 | 0.558 | 0.783 | 0.485 | 0.612 |
| Latvia | 0.296 | 0.261 | 0.121 | 0.382 | 0.346 | 0.187 | 0.540 | 0.431 | 0.884 | 0.395 | 0.876 | 0.282 | 0.663 |
| Bulgaria | 0.254 | 0.195 | 0.059 | 0.328 | 0.258 | 0.092 | 0.582 | 0.376 | 0.872 | 0.299 | 0.861 | 0.142 | 0.650 |
| Romania | 0.241 | 0.178 | 0.059 | 0.310 | 0.236 | 0.091 | 0.557 | 0.386 | 0.803 | 0.306 | 0.772 | 0.168 | 0.542 |
| $U(I^*_{MPA})$ | 0.776 | 0.756 | 0.647 | | | | | | | | | | |
| Average | 0.516 | 0.472 | 0.251 | 0.664 | 0.625 | 0.387 | 0.864 | 0.774 | 0.854 | 0.747 | 0.833 | 0.647 | 0.602 |
| Median | 0.497 | 0.469 | 0.240 | 0.640 | 0.622 | 0.370 | 0.899 | 0.772 | 0.859 | 0.758 | 0.848 | 0.642 | 0.643 |
| Minimum | 0.241 | 0.178 | 0.059 | 0.310 | 0.236 | 0.091 | 0.540 | 0.376 | 0.755 | 0.299 | 0.672 | 0.142 | 0.375 |
| Maximum | 0.749 | 0.727 | 0.495 | 0.954 | 0.950 | 0.749 | 0.999 | 0.998 | 0.956 | 0.998 | 0.953 | 0.997 | 0.751 |
| Stand. Dev. | 0.145 | 0.151 | 0.114 | 0.186 | 0.198 | 0.174 | 0.145 | 0.195 | 0.050 | 0.215 | 0.066 | 0.262 | 0.109 |

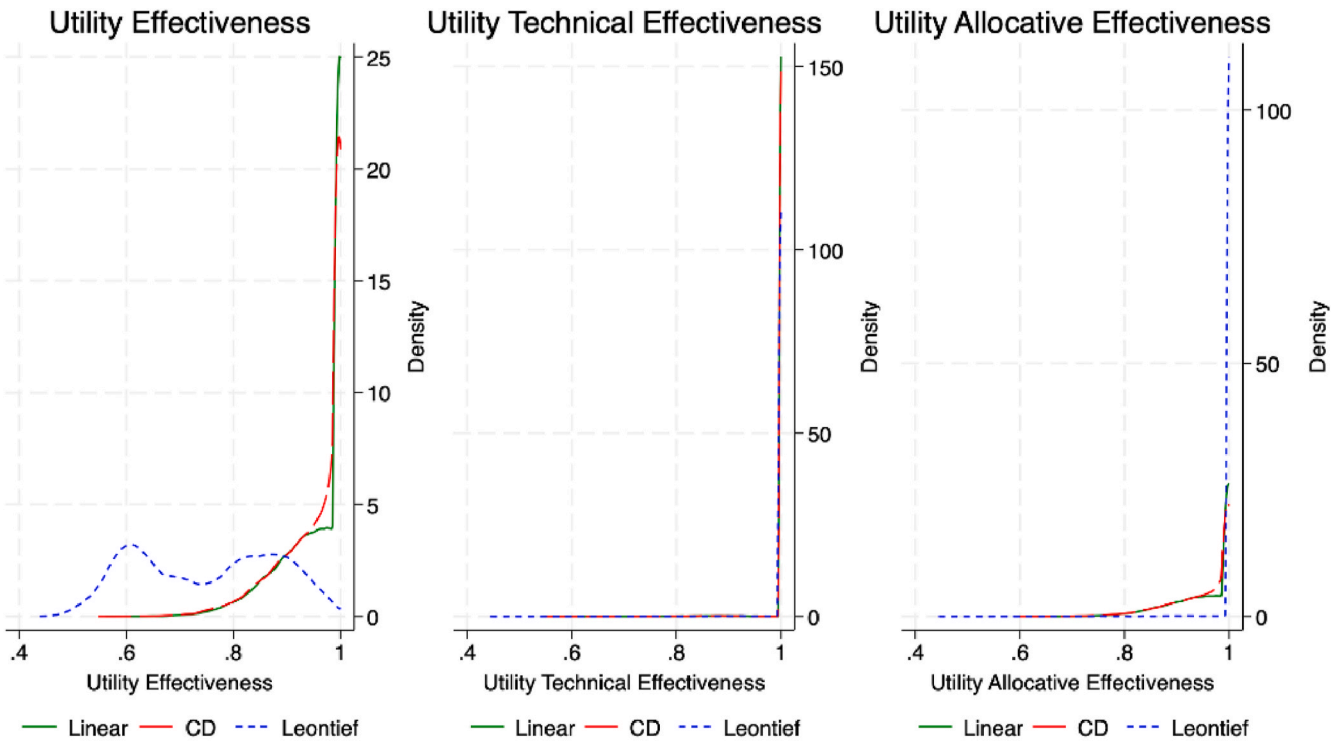
effectiveness, with values of 0.954 and 0.310 for the linear (arithmetic mean, $\overline{UE}^{LR}(I)$) specification, 0.950 and 0.236 for the Cobb-Douglas (geometric mean, $\overline{UE}^{CD}(I)$) specification, and 0.749 and 0.091 for the Leontief (fixed proportions, $\overline{UE}^{LT}(I)$) specification. Considering utility effectiveness based on the conventional linear function, $\overline{UE}^{LR}(I)$, Romania could increase its effectiveness on average by 222.6% (=(1/0.310−1)×100) by expanding its innovation indicators, thereby reaching the maximum observed utility. Alternatively, Denmark can barely increase its performance by 4.8% (=(1/0.954−1)×100) on average.

As for the sources of utility ineffectiveness, we report in Table 3 its decomposition against the strongly effective frontier (3), corresponding to expression (14). As expected, technical utility ineffectiveness is greater than its allocative counterpart for the case of the linear and Cobb-Douglas utility functions. Nevertheless, the average values are not that far apart since under the linear specification $\overline{UTE}_k^S|_{LR} = 0.774$ and $\overline{UAE}_k^S|_{LR} = 0.854$, while for the Cobb-Douglas $\overline{UTE}_k^S|_{CD} = 0.747$ and $\overline{UAE}_k^S|_{CD} = 0.883$. The sources of ineffectiveness reverses under fixed proportions, given the unbalance distributions of the indicators of national innovations systems with respect to the equal values that imposes the Leontief specification: $\overline{UTE}_k^S|_{LT} = 0.833$ and $\overline{UAE}_k^S|_{LT} = 0.647$. Hence, we observe that: 1) the lower the substitutability between indicators, the greater the gap between technical and allocative effectiveness; and 2) the choice of the Leontief function, which bans any substitutability, aggravates allocative inefficiencies and makes them the main source of ineffectiveness.

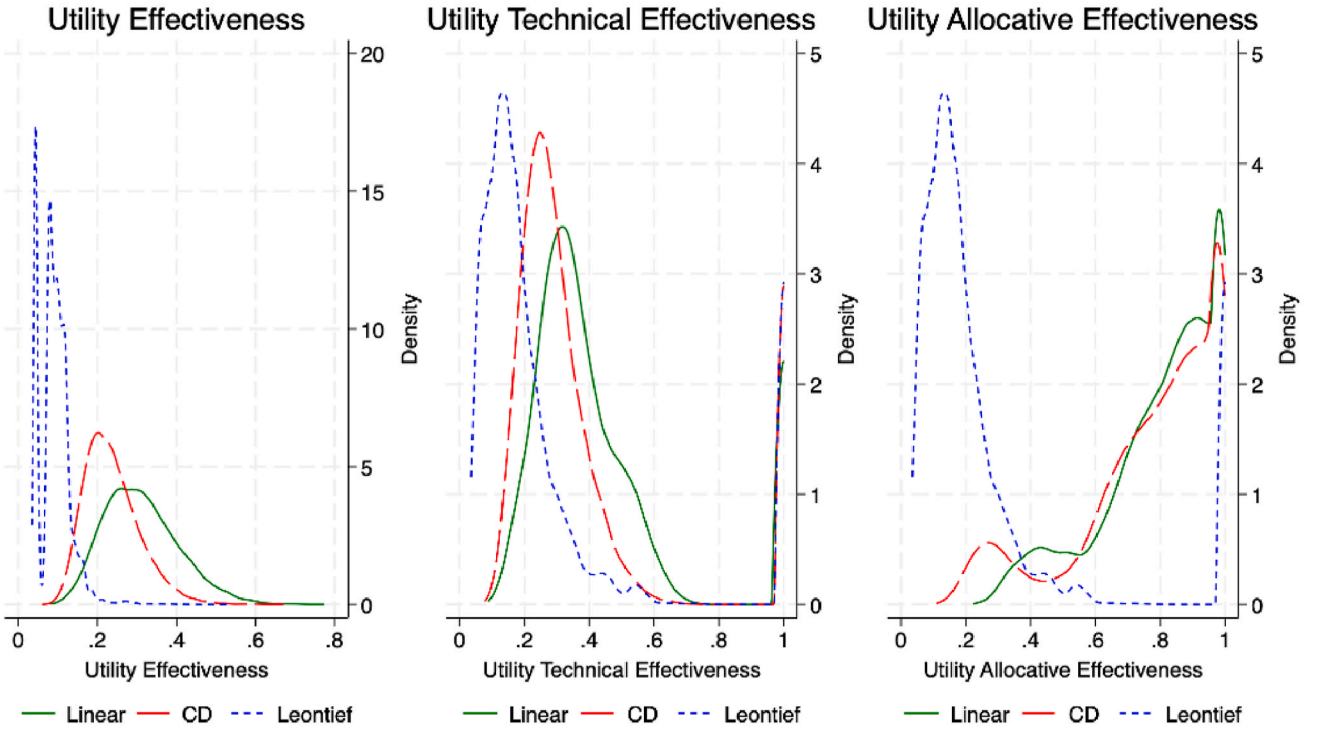
The choice of utility function thus plays a crucial role in shaping performance evaluation. While the BoD model evaluates countries based on their ability to reach the frontier under the most favorable weights, it does not incorporate the preferences of the DM regarding the relative importance and substitutability of indicators. The proposed utility effectiveness framework complements this perspective by evaluating performance under an explicit utility function, allowing us to assess not only whether a country is efficient, but whether it performs well according to a given preference structure. Therefore, beyond measuring relative performance, the proposed framework enriches the standard BoD analysis by introducing a preference-based benchmark that allows us to evaluate how well countries perform under explicit normative criteria defined by the DM.

Fig. 2 depicts the kernel density functions of these two countries for the 50,000 samples under the three specifications. While the density is higher around one for Denmark—with the exception of the fixed proportions specifications that is bimodal, this is not the case for Romania, which is clearly ineffective when reaching innovation goals under any specification. In the case of Romania, the probability density for the score accumulates almost entirely in the interval [0.05, 0.60]. In contrast, Denmark accumulates almost all its density in the interval [0.08, 1.00]. The consideration of the fixed proportions specifications with the same weights adds more variability to the results because it imposes an equally balanced combination of indicators that is not normally matched by the innovation systems.

The utility effectiveness scores reveal a familiar pattern seen in BoD analyses that rely on socioeconomic indicators: the ranking of countries is positively correlated with their value. This means that nations with




**Denmark**




**Romania**

**Fig. 2.** Utility effectiveness, Utility technical effectiveness and Utility allocative effectiveness (strongly effective frontier, $F^S(T)$): Denmark and Romania.

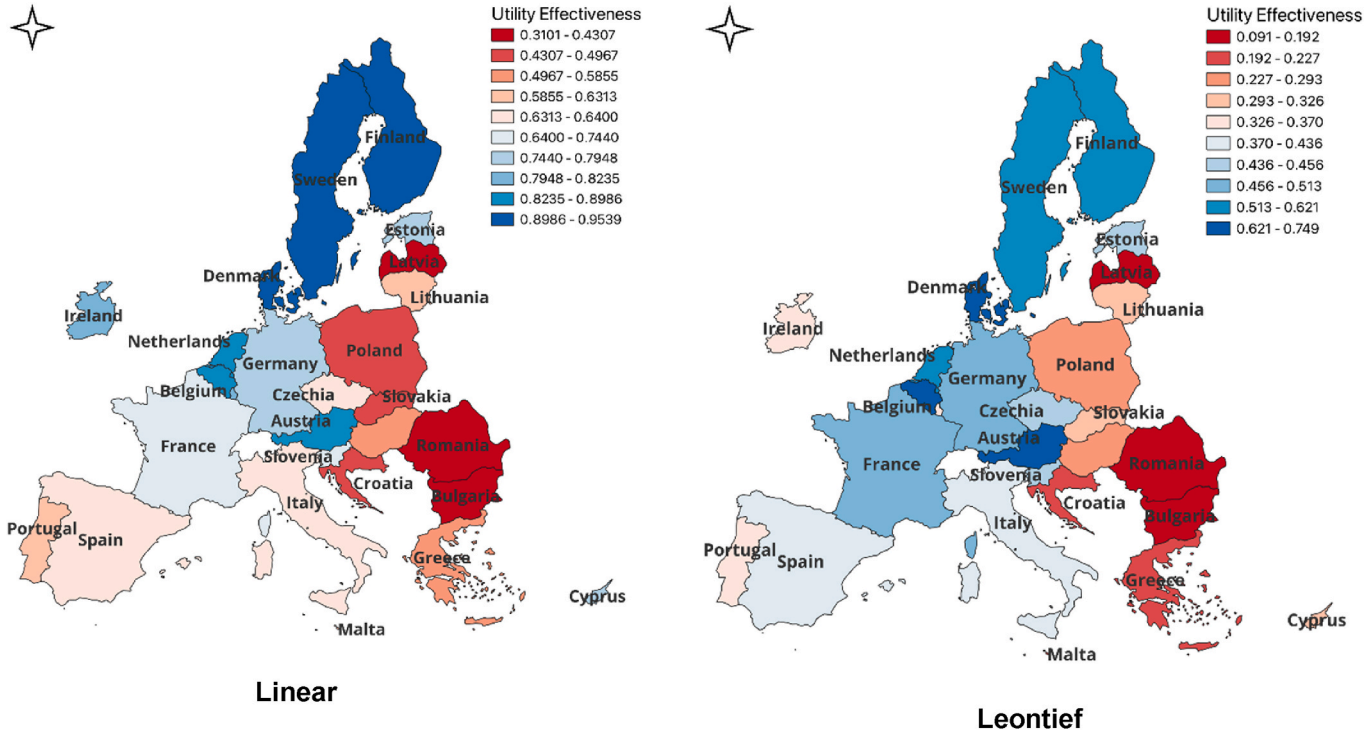



**Fig. 3.** Utility effectiveness —eq. (6). Linear $\overline{UE}^{LR}$ (left) and Leontief, $\overline{UE}^{LR}$ (right) specifications.
Note: EuroGeographics for the administrative boundaries.

well-developed and cohesive innovation systems consistently occupy the top positions, while those with weaker and fragmented systems fall behind. A clear geographical divide is evident in Fig. 3, separating northern from southern EU countries, and western from eastern ones. Scandinavian, Northern European, and Benelux nations stand out as top performers, whereas southeastern EU countries show the lowest performance levels. Given the high correlation between $\overline{UE}^{LR}$ and $\overline{UE}^{CD}$, we show the values of the former and $\overline{UE}^{LT}$. The color scale corresponds to the ten deciles of the distributions of utility effectiveness. We see that the difference in the values is large, although most of the countries belong generally to the same decile (i.e., they are depicted in the same color). Even if some countries change position from $\overline{UE}^{LR}$ to $\overline{UE}^{LT}$, they mainly move to neighboring deciles and, therefore, their color does not change but only the intensity.

The maps illustrate that correlations between the utility effectiveness specifications is also large, with $\rho(\overline{UE}^{LR}, \overline{UE}^{CD}) = 0.992$, $\rho(\overline{UE}^{LR}, \overline{UE}^{LT}) = 0.886$, and $\rho(\overline{UE}^{CD}, \overline{UE}^{LT}) = 0.933$. Besides Pearson's correlations, we also study ranking distributions using Spearman's statistic. Table 4 shows that the Spearman's correlations between the same elements of utility effectiveness, either overall, technical or allocative—presented in the cells shaded in grey—are generally above 0.9 and always statistically significant. The only exception is the correlation between the linear and the Leontief specifications of utility effectiveness $\rho_s(\overline{UE}^{CD}, \overline{UE}^{LT}) = 0.8962$, and specially, as previously discussed, the rankings of allocative effectiveness between the Leontief specifications and the other two: $\rho_s(\overline{UAE}_k^S|_{LR}, \overline{UAE}_k^S|_{LT}) = 0.6184$ and $\rho_s(\overline{UAE}_k^S|_{CD}, \overline{UAE}_k^S|_{LT}) = 0.7448$.

This shows that the ranking of countries according to pairwise utility effectiveness is quite stable, particularly regarding technical effectiveness (always above 0.97). As for Spearman's correlation between different utility effectiveness, results show that they tend to be quite high (above 0.9) and significant between all specifications when comparing overall and technical effectiveness, but greatly decrease when comparing overall and allocative effectiveness, with correlations below 0.5 and around zero for the Leontief specification, without being statistically significant in all cases. The low correlation between overall and allocative utility effectiveness in the Leontief specifications shows that the former is driven by the technical component, and that the disparities between the proportions of the observed innovation indicators and the balanced ones represented by equal weights (i.e., allocative ineffectiveness) is, as expected, completely unrelated with technical (quantity) ineffectiveness.

The decomposition of utility effectiveness into technical and allocative components is central to understanding the added value of the proposed approach. The BoD framework focuses exclusively on technical effectiveness, identifying whether a country can reach the frontier under the most favorable multipliers scheme (see Section 3.1). However, it does not assess whether the resulting combination of indicators is aligned with the particular preference structure of the DM. In contrast, our results show that countries may appear effective from a BoD perspective (technical effectiveness), yet still exhibit substantial utility losses due to an inadequate balance of indicators (allocative ineffectiveness), a dimension that is entirely absent in standard BoD analysis. This distinction is particularly relevant for policy purposes, as it differentiates between improving the overall level of indicators and correcting their internal composition.

Finally, besides the compatibility in the ranking of countries, we study whether the distributions of utility effectiveness are different between specifications. We apply Simar and Zelenyuk's [61] test—who adapt Li's [62] test to truncated data like BoD scores starting at one—to these distributions and find that they are the same for the linear and Cobb-Douglas cases, but they are different between these two and the Leontief. The results, reported in Table 5, are consistent with those reported for the Spearman correlation confirming the robustness of the attained results. In this case, utility effectiveness distributions tend to be different (equal) when their correlation is low (high). Particularly, as before, the distributions of the overall and allocative utility effectiveness under the Leontief distribution are different from their linear and

**Table 4**
Spearman's correlation matrix between utility ineffectiveness (strongly effective frontier, $F^S(T)$), EIS 2024.

| | | Utility Effectiveness: $\overline{UE}_k$ | | | Utility Technical Effectiveness: $\overline{UTE}_k^S$ | | | Utility Allocative Effectiveness: $\overline{UAE}_k^S$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Linear *LR* | Cobb-Douglas *CD* | Leontief *LT* | Linear *LR* | Cobb-Douglas *CD* | Leontief *LT* | Linear *LR* | Cobb-Douglas *CD* | Leontief *LT* |
| Utility Effectiveness: | Linear: *LR* | 1*** | | | | | | | | |
| $UE_k$ | Cobb-D.: CD | 0.9896*** | 1*** | | | | | | | |
| | Leontief: *LT* | 0.8962*** | 0.9316*** | 1*** | | | | | | |
| Utility Technical Effectiveness: | Linear: *LR* | 0.9799*** | 0.9676*** | 0.8498*** | 1*** | | | | | |
| $UTE_k^S$ | Cobb-D.: CD | 0.9829*** | 0.9707*** | 0.8541*** | 0.9994*** | 1*** | | | | |
| | Leontief: *LT* | 0.9853*** | 0.9823*** | 0.9115*** | 0.9786*** | 0.9799*** | 1*** | | | |
| Utility Allocative Effectiveness. | Linear: *LR* | 0.4731** | 0.4902** | 0.5244*** | 0.3791* | 0.3864** | 0.4292** | 1*** | | |
| $UAE_k^S$ | Cobb-D.: CD | 0.3803* | 0.4103** | 0.5128*** | 0.2772 ns | 0.2857 ns | 0.3443* | 0.9658*** | 1*** | |
| | Leontief: *LT* | −0.0012 ns | 0.0488 ns | 0.2918 ns | −0.1319 ns | −0.1209 ns | −0.0165 | 0.6184*** | 0.7448*** | 1*** |

Notes: ***p < 0.01; **p < 0.05; *p < 0.1; ns Non-significant at 10% level.

**Table 5**
Li tests for inequality of distributions of utility ineffectiveness (strongly effective frontier, $F^S(T)$), EIS 2024.

| | | Utility Effectiveness: $\overline{UE}_k$ | | | Utility Technical Effectiveness: $\overline{UTE}_k^S$ | | | Utility Allocative Effectiveness: $\overline{UAE}_k^S$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Linear *LR* | Cobb-Douglas *CD* | Leontief *LT* | Linear *LR* | Cobb-Douglas *CD* | Leontief *LT* | Linear *LR* | Cobb-Douglas *CD* | Leontief *LT* |
| Utility Effectiveness: | Linear: *LR* | – | | | | | | | | |
| $UE_k$ | Cobb-D.: CD | 0.0129 ns | – | | | | | | | |
| | Leontief: *LT* | 4.7152*** | 3.1287*** | – | | | | | | |
| Utility Technical Effectiveness: | Linear: *LR* | 1.1339 ns | 1.6003** | 8.8583*** | – | | | | | |
| $UTE_k^S$ | Cobb-D.: CD | 0.9707 ns | 1.292* | 7.5001*** | 0.005 ns | – | | | | |
| | Leontief: *LT* | 0.6351 ns | 0.6631 | 3.7835*** | 0.4105 ns | 0.221 ns | – | | | |
| Utility Allocative Effectiveness. | Linear: *LR* | 5.3739*** | 5.7026*** | 12.3664*** | 5.973*** | 6.8195*** | 5.7144*** | – | | |
| $UAE_k^S$ (T1) | Cobb-D.: CD | 4.862*** | 5.3493*** | 13.5499*** | 5.4495*** | 5.9376*** | 5.698*** | −0.047ns | – | |
| | Leontief: *LT* | 2.1167** | 1.8245** | 6.0598*** | 6.6227*** | 5.7318*** | 4.1875*** | 13.6779*** | 12.9711*** | – |

Notes: ***p < 0.01; **p < 0.05; *p < 0.1; ns Non-significant at 10% level. Asterisks denote statistically significant differences between models at the stated percent level.

**Table 6**

Benchmark peers. Utility Technical Effectiveness. $\widehat{\lambda}^{*}_{ki} > 0$ (program (7)) in %.

| Country | Austria | Belgium | Bulgaria | Cyprus | Czechia | Germany | Denmark | Estonia | Greece | Spain | Finland | France | Croatia | Hungary | Ireland | Italy | Lithuania | Luxemb. | Latvia | Malta | Netherl. | Poland | Portugal | Romania | Sweden | Slovenia | Slovakia |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Austria | 75.6 | 6.9 | 0.0 | 4.4 | 0.8 | 5.4 | 14.3 | 2.1 | 1.5 | 0.3 | 5.3 | 0.4 | 0.0 | 0.2 | 4.6 | 0.7 | 0.1 | 3.6 | 0.0 | 3.9 | 4.5 | 0.0 | 0.1 | 0.0 | 9.8 | 1.0 | 0.2 |
| Belgium | 0.8 | 83.8 | 0.0 | 3.1 | 0.9 | 2.4 | 7.1 | 1.1 | 1.7 | 0.6 | 5.7 | 0.4 | 0.0 | 0.1 | 3.9 | 0.2 | 0.6 | 1.8 | 0.0 | 0.7 | 3.2 | 0.0 | 0.0 | 0.0 | 9.2 | 0.3 | 0.1 |
| Bulgaria | 17.6 | 3.0 | 2.0 | 14.7 | 6.3 | 16.1 | 32.2 | 13.8 | 2.0 | 3.5 | 14.8 | 1.3 | 0.0 | 1.2 | 16.0 | 0.8 | 6.4 | 13.5 | 0.0 | 11.4 | 2.3 | 1.0 | 0.9 | 1.6 | 19.6 | 1.2 | 2.4 |
| Cyprus | 0.6 | 1.9 | 0.0 | 86.0 | 0.6 | 1.2 | 6.1 | 0.6 | 0.9 | 0.7 | 3.0 | 0.0 | 0.0 | 0.0 | 2.6 | 0.2 | 1.2 | 1.6 | 0.0 | 0.6 | 1.8 | 0.0 | 0.2 | 0.2 | 4.4 | 0.1 | 0.1 |
| Czechia | 12.4 | 17.0 | 0.0 | 4.8 | 40.9 | 13.4 | 23.0 | 4.4 | 5.5 | 3.4 | 19.6 | 3.8 | 0.1 | 0.1 | 16.0 | 1.4 | 0.3 | 5.8 | 0.0 | 8.6 | 9.2 | 0.0 | 0.3 | 0.0 | 16.6 | 0.3 | 0.2 |
| Germany | 2.1 | 6.4 | 0.0 | 2.8 | 0.2 | 80.1 | 10.9 | 1.3 | 0.7 | 0.4 | 4.4 | 0.3 | 0.0 | 0.0 | 4.4 | 0.8 | 0.4 | 3.2 | 0.0 | 2.2 | 2.2 | 0.0 | 0.1 | 0.0 | 4.9 | 0.1 | 0.0 |
| Denmark | 0.2 | 0.6 | 0.0 | 0.4 | 0.1 | 0.2 | 98.3 | 0.2 | 0.1 | 0.0 | 0.8 | 0.1 | 0.0 | 0.0 | 0.7 | 0.0 | 0.0 | 0.4 | 0.0 | 0.1 | 0.6 | 0.0 | 0.0 | 0.0 | 1.1 | 0.0 | 0.0 |
| Estonia | 6.4 | 8.2 | 0.0 | 7.6 | 0.3 | 3.0 | 19.3 | 56.3 | 2.7 | 0.1 | 16.6 | 2.6 | 0.0 | 0.0 | 4.0 | 1.1 | 1.3 | 6.6 | 0.0 | 1.5 | 9.1 | 0.1 | 0.0 | 0.0 | 13.8 | 0.2 | 0.0 |
| Greece | 5.4 | 19.2 | 0.0 | 9.4 | 3.4 | 7.5 | 14.2 | 7.3 | 46.4 | 3.3 | 6.0 | 2.2 | 0.0 | 0.1 | 12.2 | 0.9 | 3.6 | 7.2 | 0.0 | 2.9 | 7.2 | 0.0 | 0.3 | 0.0 | 8.6 | 0.0 | 0.0 |
| Spain | 8.0 | 8.1 | 0.0 | 6.5 | 4.7 | 9.9 | 27.3 | 8.3 | 0.1 | 34.3 | 9.6 | 9.7 | 0.0 | 0.7 | 14.1 | 2.2 | 6.1 | 13.9 | 0.0 | 7.2 | 20.6 | 0.0 | 1.6 | 1.2 | 12.3 | 1.0 | 0.5 |
| Finland | 0.7 | 1.2 | 0.0 | 1.5 | 0.1 | 0.8 | 5.5 | 0.5 | 0.4 | 0.1 | 91.8 | 0.3 | 0.0 | 0.0 | 1.1 | 0.2 | 0.1 | 1.2 | 0.0 | 0.2 | 1.1 | 0.0 | 0.0 | 0.0 | 4.8 | 0.0 | 0.0 |
| France | 7.3 | 13.9 | 0.0 | 7.2 | 1.8 | 16.0 | 23.2 | 1.8 | 1.8 | 0.8 | 9.3 | 46.0 | 0.0 | 0.7 | 17.2 | 2.3 | 1.0 | 15.2 | 0.0 | 6.0 | 12.8 | 0.0 | 0.2 | 0.2 | 21.0 | 1.8 | 0.6 |
| Croatia | 5.9 | 26.5 | 0.0 | 19.9 | 3.6 | 17.7 | 28.7 | 10.1 | 20.5 | 3.9 | 29.7 | 6.6 | 1.9 | 0.2 | 14.6 | 5.1 | 4.2 | 8.9 | 0.0 | 4.4 | 9.5 | 0.0 | 0.0 | 0.0 | 24.2 | 3.0 | 0.4 |
| Hungary | 10.9 | 11.2 | 0.0 | 6.8 | 8.1 | 10.8 | 28.9 | 8.3 | 0.3 | 2.6 | 14.0 | 10.3 | 0.0 | 22.4 | 17.5 | 0.4 | 7.6 | 13.1 | 0.0 | 14.4 | 17.8 | 0.0 | 4.6 | 0.3 | 12.8 | 0.1 | 0.9 |
| Ireland | 0.9 | 2.8 | 0.0 | 1.9 | 0.1 | 0.8 | 3.3 | 0.5 | 0.2 | 0.0 | 2.5 | 0.4 | 0.0 | 0.1 | 87.9 | 0.4 | 0.4 | 4.1 | 0.0 | 0.8 | 5.0 | 0.0 | 0.1 | 0.0 | 3.3 | 0.2 | 0.0 |
| Italy | 12.9 | 15.2 | 0.0 | 11.5 | 4.9 | 15.4 | 21.2 | 6.3 | 10.0 | 1.7 | 11.7 | 1.4 | 0.0 | 0.8 | 11.2 | 34.5 | 1.7 | 5.5 | 0.0 | 6.3 | 13.1 | 0.0 | 0.2 | 0.0 | 20.0 | 2.9 | 0.9 |
| Lithuania | 5.9 | 11.1 | 0.0 | 12.8 | 0.4 | 6.5 | 17.6 | 7.9 | 3.4 | 2.3 | 12.5 | 3.6 | 0.1 | 0.0 | 11.4 | 1.1 | 46.5 | 7.0 | 0.0 | 7.3 | 4.0 | 0.0 | 0.7 | 0.4 | 14.3 | 0.4 | 0.1 |
| Luxembourg | 0.7 | 0.3 | 0.0 | 2.3 | 0.1 | 0.7 | 7.7 | 0.9 | 0.1 | 0.2 | 1.4 | 0.1 | 0.0 | 0.0 | 1.5 | 0.3 | 0.5 | 88.0 | 0.0 | 1.5 | 2.1 | 0.0 | 0.2 | 0.3 | 1.5 | 0.1 | 0.0 |
| Latvia | 7.9 | 5.4 | 0.1 | 27.6 | 4.8 | 15.7 | 35.7 | 18.1 | 0.7 | 2.2 | 20.1 | 1.3 | 0.0 | 0.4 | 24.7 | 1.6 | 12.1 | 31.1 | 0.0 | 11.3 | 12.6 | 0.2 | 0.2 | 0.4 | 27.7 | 1.5 | 0.7 |
| Malta | 2.7 | 1.8 | 0.0 | 3.2 | 1.9 | 3.5 | 12.6 | 1.6 | 0.1 | 1.9 | 5.9 | 0.2 | 0.0 | 1.1 | 6.1 | 1.1 | 2.2 | 5.7 | 0.0 | 67.1 | 7.2 | 0.1 | 0.6 | 1.3 | 7.9 | 1.3 | 1.2 |
| Netherlands | 2.2 | 1.6 | 0.0 | 3.1 | 0.1 | 1.7 | 8.5 | 1.1 | 0.1 | 0.1 | 6.2 | 0.9 | 0.0 | 0.1 | 3.8 | 0.2 | 0.7 | 3.9 | 0.0 | 1.3 | 82.2 | 0.1 | 0.1 | 0.1 | 10.6 | 0.3 | 0.0 |
| Poland | 22.0 | 11.3 | 0.0 | 14.2 | 8.6 | 16.7 | 39.8 | 8.7 | 0.3 | 3.9 | 14.1 | 4.9 | 0.0 | 3.9 | 16.2 | 0.3 | 10.4 | 11.0 | 0.0 | 15.1 | 10.1 | 3.5 | 3.7 | 1.0 | 21.2 | 4.6 | 3.5 |
| Portugal | 13.4 | 16.7 | 0.0 | 15.6 | 3.2 | 12.6 | 31.4 | 10.3 | 8.0 | 4.1 | 21.4 | 7.1 | 0.0 | 0.2 | 12.5 | 2.8 | 2.3 | 12.8 | 0.0 | 7.7 | 13.8 | 0.1 | 23.0 | 3.9 | 20.2 | 2.4 | 0.3 |
| Romania | 5.0 | 7.3 | 0.1 | 16.4 | 5.7 | 14.3 | 27.7 | 3.9 | 1.5 | 0.8 | 8.8 | 1.8 | 0.0 | 6.6 | 21.5 | 0.6 | 3.6 | 10.1 | 0.0 | 9.6 | 7.2 | 0.0 | 0.1 | 16.4 | 9.3 | 2.0 | 10.5 |
| Sweden | 0.2 | 0.4 | 0.0 | 0.5 | 0.1 | 0.2 | 1.7 | 0.2 | 0.1 | 0.0 | 0.6 | 0.1 | 0.0 | 0.0 | 0.6 | 0.0 | 0.0 | 0.5 | 0.0 | 0.2 | 0.5 | 0.0 | 0.0 | 0.0 | 97.9 | 0.0 | 0.0 |
| Slovenia | 15.4 | 22.9 | 0.0 | 17.6 | 1.3 | 14.0 | 25.9 | 7.8 | 12.6 | 2.5 | 20.4 | 3.8 | 0.0 | 1.4 | 10.8 | 2.9 | 1.4 | 9.8 | 0.0 | 8.6 | 12.2 | 0.1 | 1.8 | 0.0 | 29.1 | 23.4 | 1.5 |
| Slovakia | 8.5 | 8.2 | 0.0 | 7.7 | 8.7 | 14.3 | 26.9 | 9.0 | 3.6 | 9.0 | 13.9 | 3.6 | 0.0 | 0.1 | 18.8 | 1.5 | 6.9 | 12.6 | 0.0 | 12.8 | 10.0 | 0.0 | 1.2 | 0.1 | 14.3 | 0.1 | 22.8 |
| Average | 6.77 | 8.82 | 0.01 | 8.60 | 2.72 | 8.49 | 19.25 | 5.23 | 3.03 | 1.86 | 10.71 | 2.58 | 0.02 | 0.68 | 10.31 | 1.12 | 2.88 | 8.09 | 0.00 | 5.65 | 7.67 | 0.07 | 0.66 | 0.43 | 13.17 | 0.96 | 0.93 |
| Median | 5.93 | 7.68 | 0.01 | 7.00 | 1.60 | 8.70 | 20.25 | 4.16 | 1.20 | 1.26 | 9.42 | 1.37 | 0.01 | 0.09 | 11.31 | 0.79 | 1.32 | 6.82 | 0.00 | 5.19 | 7.21 | 0.02 | 0.16 | 0.06 | 12.54 | 0.31 | 0.18 |
| Minimum | 0.15 | 0.33 | 0.00 | 0.37 | 0.05 | 0.19 | 1.65 | 0.15 | 0.05 | 0.01 | 0.64 | 0.01 | 0.00 | 0.00 | 0.57 | 0.03 | 0.03 | 0.35 | 0.00 | 0.13 | 0.54 | 0.00 | 0.00 | 0.00 | 1.05 | 0.01 | 0.00 |
| Maximum | 21.95 | 26.55 | 0.13 | 27.63 | 8.66 | 17.71 | 39.80 | 18.08 | 20.52 | 9.01 | 29.71 | 10.28 | 0.07 | 6.60 | 24.73 | 5.11 | 12.06 | 31.13 | 0.00 | 15.08 | 20.57 | 1.00 | 4.65 | 3.86 | 29.06 | 4.59 | 10.50 |
| Stand. Dev. | 5.97 | 7.33 | 0.03 | 6.96 | 2.90 | 6.43 | 10.87 | 4.80 | 4.84 | 2.05 | 7.49 | 2.97 | 0.02 | 1.45 | 7.12 | 1.16 | 3.37 | 6.55 | 0.00 | 4.77 | 5.46 | 0.20 | 1.15 | 0.84 | 7.89 | 1.19 | 2.12 |

Note: For each evaluated country (rows), the columns show the number of times—in percentage terms over the number of random samples where the country 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, implying that it is its own benchmark. More intense color shades indicate bigger percentages (green) and lower percentages (red). Values in the diagonal is excluded from the summary statistics of each country.

Cobb-Douglas counterparts at the 1% level of confidence, while they are equal regarding technical effectiveness. As for the equality of distributions between the linear and Cobb-Douglas specifications, it cannot be rejected at 1% level of confidence, indicating that, even if the results could be different for individual countries, one could use any of the two specifications to obtain results that, in general for the whole sample, are not statistically different.

### 4.3.2. *Frontier peers and slacks*

It is relevant for stakeholders to identify the benchmark country(ies) that maximize the utility when solving program (5), necessary to determine utility effectiveness, $UE_k$, as well as all those that serve as reference peers of the radial projections on the utility frontier when solving program (7), required in turn to calculate the utility technical effectiveness $UTE_k$. We focus first on these latter peers that Pareto dominate ineffective countries from a technical perspective. As discussed in Section 3, it is possible to determine the countries that serve as reference when constructing the linear combinations that conform the frontier peers by looking at the optimal values of their associated multipliers. A positive value of the multiplier in program (7), i.e., $\widehat{\lambda}^{*}_{k} > 0$ indicates that country $k$ serves as reference target when measuring technical effectiveness.

**Table 7**

Average value of slacks, 50,000 samples. Utility Technical Effectiveness (program (8)).

| Country | 1.1.1 | 1.1.2 | 1.1.3 | 1.2.1 | 1.2.2 | 1.2.3 | 1.3.1 | 1.3.2 | 2.1.1 | 2.1.2 | 2.1.3 | 2.2.1 | 2.2.2 | 2.2.3 | 2.3.1 | 2.3.2 | 3.1.1 | 3.1.2 | 3.2.1 | 3.2.2 | 3.2.3 | 3.3.1 | 3.3.2 | 3.3.3 | 4.1.1 | 4.1.2 | 4.2.1 | 4.2.2 | 4.2.3 | 4.3.1 | 4.3.2 | 4.3.3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Austria | 0.10 | 0.16 | 0.22 | 0.17 | 0.07 | 0.07 | 0.36 | 0.11 | 0.06 | 0.28 | 0.00 | 0.02 | 0.13 | 0.16 | 0.29 | 0.13 | 0.06 | 0.11 | 0.18 | 0.10 | 0.17 | 0.08 | 0.08 | 0.00 | 0.14 | 0.07 | 0.07 | 0.31 | 0.08 | 0.15 | 0.02 | 0.20 |
| Belgium | 0.08 | 0.09 | 0.45 | 0.16 | 0.03 | 0.15 | 0.12 | 0.20 | 0.07 | 0.11 | 0.00 | 0.00 | 0.05 | 0.00 | 0.05 | 0.20 | 0.04 | 0.00 | 0.09 | 0.18 | 0.22 | 0.19 | 0.12 | 0.15 | 0.16 | 0.00 | 0.06 | 0.05 | 0.08 | 0.02 | 0.06 | 0.19 |
| Bulgaria | 0.23 | 0.21 | 0.65 | 0.54 | 0.42 | 0.33 | 0.17 | 0.51 | 0.48 | 0.25 | 0.38 | 0.39 | 0.16 | 0.46 | 0.46 | 0.12 | 0.24 | 0.48 | 0.33 | 0.52 | 0.52 | 0.37 | 0.04 | 0.05 | 0.19 | 0.24 | 0.22 | 0.20 | 0.18 | 0.35 | 0.35 | 0.28 |
| Cyprus | 0.46 | 0.00 | 0.44 | 0.00 | 0.06 | 0.17 | 0.08 | 0.25 | 0.52 | 0.44 | 0.32 | 0.59 | 0.09 | 0.52 | 0.12 | 0.17 | 0.04 | 0.00 | 0.00 | 0.20 | 0.02 | 0.45 | 0.00 | 0.17 | 0.07 | 0.08 | 0.12 | 0.09 | 0.12 | 0.22 | 0.31 | 0.31 |
| Czechia | 0.08 | 0.33 | 0.33 | 0.28 | 0.34 | 0.15 | 0.26 | 0.06 | 0.14 | 0.13 | 0.18 | 0.17 | 0.00 | 0.09 | 0.14 | 0.19 | 0.07 | 0.11 | 0.18 | 0.33 | 0.34 | 0.39 | 0.11 | 0.16 | 0.14 | 0.07 | 0.00 | 0.19 | 0.10 | 0.17 | 0.03 | 0.24 |
| Germany | 0.05 | 0.30 | 0.44 | 0.45 | 0.07 | 0.19 | 0.15 | 0.33 | 0.08 | 0.17 | 0.28 | 0.04 | 0.04 | 0.07 | 0.13 | 0.18 | 0.12 | 0.13 | 0.29 | 0.34 | 0.09 | 0.04 | 0.10 | 0.10 | 0.16 | 0.03 | 0.00 | 0.06 | 0.10 | 0.11 | 0.02 | 0.22 |
| Denmark | 0.04 | 0.14 | 0.00 | 0.00 | 0.01 | 0.07 | 0.00 | 0.10 | 0.00 | 0.00 | 0.12 | 0.12 | 0.00 | 0.12 | 0.06 | 0.16 | 0.07 | 0.08 | 0.29 | 0.00 | 0.00 | 0.02 | 0.06 | 0.00 | 0.19 | 0.17 | 0.08 | 0.04 | 0.10 | 0.24 | 0.00 | 0.00 |
| Estonia | 0.10 | 0.15 | 0.10 | 0.21 | 0.08 | 0.10 | 0.35 | 0.18 | 0.16 | 0.00 | 0.24 | 0.35 | 0.00 | 0.24 | 0.35 | 0.09 | 0.08 | 0.12 | 0.13 | 0.35 | 0.11 | 0.41 | 0.00 | 0.08 | 0.13 | 0.03 | 0.11 | 0.10 | 0.17 | 0.29 | 0.12 | 0.35 |
| Greece | 0.11 | 0.16 | 0.47 | 0.27 | 0.12 | 0.53 | 0.43 | 0.23 | 0.12 | 0.30 | 0.36 | 0.38 | 0.05 | 0.33 | 0.36 | 0.50 | 0.00 | 0.03 | 0.28 | 0.24 | 0.26 | 0.31 | 0.16 | 0.26 | 0.23 | 0.00 | 0.42 | 0.16 | 0.00 | 0.29 | 0.12 | 0.16 |
| Spain | 0.12 | 0.13 | 0.21 | 0.34 | 0.13 | 0.25 | 0.00 | 0.06 | 0.20 | 0.14 | 0.11 | 0.28 | 0.09 | 0.29 | 0.18 | 0.21 | 0.22 | 0.29 | 0.27 | 0.38 | 0.16 | 0.22 | 0.09 | 0.15 | 0.23 | 0.35 | 0.09 | 0.41 | 0.00 | 0.06 | 0.12 | 0.29 |
| Finland | 0.06 | 0.34 | 0.05 | 0.11 | 0.04 | 0.14 | 0.03 | 0.00 | 0.10 | 0.00 | 0.21 | 0.09 | 0.10 | 0.15 | 0.00 | 0.00 | 0.05 | 0.09 | 0.09 | 0.11 | 0.16 | 0.00 | 0.07 | 0.13 | 0.17 | 0.02 | 0.11 | 0.04 | 0.08 | 0.35 | 0.08 | 0.26 |
| France | 0.05 | 0.08 | 0.23 | 0.45 | 0.15 | 0.07 | 0.12 | 0.15 | 0.16 | 0.00 | 0.00 | 0.15 | 0.15 | 0.10 | 0.39 | 0.20 | 0.08 | 0.06 | 0.17 | 0.46 | 0.13 | 0.14 | 0.19 | 0.16 | 0.16 | 0.10 | 0.06 | 0.11 | 0.31 | 0.08 | 0.07 | 0.25 |
| Croatia | 0.12 | 0.21 | 0.37 | 0.33 | 0.36 | 0.34 | 0.32 | 0.20 | 0.16 | 0.02 | 0.38 | 0.32 | 0.11 | 0.52 | 0.21 | 0.19 | 0.02 | 0.11 | 0.23 | 0.25 | 0.18 | 0.39 | 0.17 | 0.24 | 0.19 | 0.11 | 0.13 | 0.59 | 0.09 | 0.15 | 0.15 | 0.37 |
| Hungary | 0.23 | 0.47 | 0.24 | 0.35 | 0.21 | 0.11 | 0.14 | 0.08 | 0.29 | 0.17 | 0.05 | 0.18 | 0.10 | 0.21 | 0.15 | 0.11 | 0.14 | 0.34 | 0.18 | 0.28 | 0.12 | 0.18 | 0.17 | 0.34 | 0.11 | 0.33 | 0.00 | 0.10 | 0.17 | 0.20 | 0.07 | 0.22 |
| Ireland | 0.13 | 0.00 | 0.33 | 0.16 | 0.07 | 0.12 | 0.13 | 0.04 | 0.57 | 0.28 | 0.15 | 0.36 | 0.16 | 0.00 | 0.17 | 0.11 | 0.12 | 0.14 | 0.12 | 0.17 | 0.09 | 0.25 | 0.24 | 0.30 | 0.00 | 0.13 | 0.04 | 0.00 | 0.00 | 0.04 | 0.00 | 0.26 |
| Italy | 0.08 | 0.47 | 0.31 | 0.34 | 0.10 | 0.34 | 0.23 | 0.24 | 0.27 | 0.27 | 0.27 | 0.36 | 0.06 | 0.22 | 0.23 | 0.23 | 0.08 | 0.05 | 0.20 | 0.30 | 0.34 | 0.20 | 0.11 | 0.01 | 0.16 | 0.12 | 0.05 | 0.21 | 0.07 | 0.00 | 0.06 | 0.21 |
| Lithuania | 0.23 | 0.02 | 0.29 | 0.44 | 0.24 | 0.41 | 0.14 | 0.18 | 0.28 | 0.04 | 0.25 | 0.41 | 0.00 | 0.24 | 0.41 | 0.17 | 0.15 | 0.13 | 0.25 | 0.48 | 0.00 | 0.39 | 0.14 | 0.21 | 0.18 | 0.08 | 0.17 | 0.40 | 0.15 | 0.26 | 0.05 | 0.27 |
| Luxembourg | 0.00 | 0.02 | 0.30 | 0.00 | 0.07 | 0.00 | 0.08 | 0.22 | 0.41 | 0.23 | 0.27 | 0.46 | 0.21 | 0.34 | 0.24 | 0.00 | 0.16 | 0.16 | 0.26 | 0.09 | 0.07 | 0.29 | 0.11 | 0.15 | 0.00 | 0.20 | 0.10 | 0.06 | 0.45 | 0.00 | 0.07 | 0.38 |
| Latvia | 0.29 | 0.07 | 0.14 | 0.38 | 0.26 | 0.27 | 0.18 | 0.19 | 0.09 | 0.17 | 0.28 | 0.36 | 0.11 | 0.48 | 0.15 | 0.14 | 0.20 | 0.15 | 0.29 | 0.29 | 0.10 | 0.14 | 0.08 | 0.09 | 0.22 | 0.27 | 0.12 | 0.04 | 0.19 | 0.25 | 0.21 | 0.16 |
| Malta | 0.49 | 0.16 | 0.23 | 0.30 | 0.17 | 0.00 | 0.02 | 0.06 | 0.54 | 0.61 | 0.41 | 0.43 | 0.11 | 0.38 | 0.09 | 0.20 | 0.30 | 0.24 | 0.41 | 0.39 | 0.19 | 0.23 | 0.00 | 0.00 | 0.11 | 0.22 | 0.03 | 0.35 | 0.32 | 0.11 | 0.00 | 0.24 |
| Netherlands | 0.20 | 0.06 | 0.02 | 0.13 | 0.00 | 0.06 | 0.07 | 0.00 | 0.15 | 0.05 | 0.07 | 0.16 | 0.21 | 0.22 | 0.11 | 0.06 | 0.11 | 0.06 | 0.21 | 0.12 | 0.10 | 0.09 | 0.08 | 0.11 | 0.07 | 0.09 | 0.05 | 0.04 | 0.23 | 0.00 | 0.05 | 0.20 |
| Poland | 0.32 | 0.07 | 0.33 | 0.47 | 0.24 | 0.54 | 0.14 | 0.21 | 0.27 | 0.33 | 0.24 | 0.24 | 0.08 | 0.31 | 0.17 | 0.15 | 0.26 | 0.26 | 0.30 | 0.49 | 0.16 | 0.36 | 0.11 | 0.00 | 0.17 | 0.28 | 0.07 | 0.19 | 0.19 | 0.21 | 0.17 | 0.27 |
| Portugal | 0.12 | 0.18 | 0.23 | 0.22 | 0.13 | 0.05 | 0.00 | 0.14 | 0.21 | 0.38 | 0.00 | 0.22 | 0.13 | 0.41 | 0.16 | 0.16 | 0.09 | 0.09 | 0.30 | 0.31 | 0.18 | 0.31 | 0.04 | 0.17 | 0.20 | 0.08 | 0.11 | 0.34 | 0.02 | 0.21 | 0.77 | 0.26 |
| Romania | 0.28 | 0.16 | 0.31 | 0.52 | 0.21 | 0.47 | 0.00 | 0.44 | 0.51 | 0.31 | 0.29 | 0.39 | 0.34 | 0.48 | 0.45 | 0.33 | 0.48 | 0.18 | 0.18 | 0.48 | 0.23 | 0.40 | 0.18 | 0.27 | 0.48 | 0.16 | 0.00 | 0.26 | 0.28 | 0.40 | 0.23 | 0.36 |
| Sweden | 0.06 | 0.12 | 0.00 | 0.10 | 0.03 | 0.09 | 0.04 | 0.09 | 0.11 | 0.04 | 0.10 | 0.00 | 0.11 | 0.07 | 0.06 | 0.00 | 0.02 | 0.10 | 0.00 | 0.12 | 0.58 | 0.00 | 0.09 | 0.16 | 0.06 | 0.04 | 0.09 | 0.05 | 0.13 | 0.22 | 0.06 | 0.29 |
| Slovenia | 0.07 | 0.18 | 0.10 | 0.15 | 0.16 | 0.16 | 0.14 | 0.28 | 0.18 | 0.41 | 0.10 | 0.12 | 0.28 | 0.36 | 0.07 | 0.28 | 0.00 | 0.15 | 0.25 | 0.14 | 0.10 | 0.19 | 0.14 | 0.12 | 0.14 | 0.09 | 0.00 | 0.35 | 0.09 | 0.15 | 0.08 | 0.16 |
| Slovakia | 0.14 | 0.16 | 0.21 | 0.30 | 0.30 | 0.28 | 0.24 | 0.20 | 0.29 | 0.30 | 0.17 | 0.28 | 0.09 | 0.27 | 0.21 | 0.12 | 0.18 | 0.22 | 0.16 | 0.33 | 0.43 | 0.28 | 0.16 | 0.23 | 0.20 | 0.23 | 0.00 | 0.20 | 0.20 | 0.13 | 0.06 | 0.26 |
| Average | 0.16 | 0.16 | 0.26 | 0.27 | 0.15 | 0.21 | 0.14 | 0.18 | 0.24 | 0.20 | 0.20 | 0.26 | 0.11 | 0.27 | 0.20 | 0.16 | 0.13 | 0.14 | 0.21 | 0.28 | 0.19 | 0.24 | 0.11 | 0.15 | 0.16 | 0.14 | 0.09 | 0.18 | 0.15 | 0.17 | 0.13 | 0.25 |
| Median | 0.12 | 0.15 | 0.27 | 0.29 | 0.13 | 0.15 | 0.13 | 0.19 | 0.19 | 0.17 | 0.22 | 0.28 | 0.10 | 0.26 | 0.17 | 0.17 | 0.10 | 0.12 | 0.22 | 0.30 | 0.16 | 0.24 | 0.11 | 0.15 | 0.16 | 0.10 | 0.08 | 0.14 | 0.12 | 0.18 | 0.07 | 0.26 |
| Minimum | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Maximum | 0.49 | 0.47 | 0.65 | 0.54 | 0.42 | 0.54 | 0.43 | 0.51 | 0.57 | 0.61 | 0.41 | 0.59 | 0.34 | 0.52 | 0.46 | 0.50 | 0.48 | 0.48 | 0.41 | 0.52 | 0.58 | 0.45 | 0.24 | 0.34 | 0.48 | 0.35 | 0.42 | 0.59 | 0.45 | 0.40 | 0.77 | 0.38 |
| Stand. Dev. | 0.12 | 0.13 | 0.16 | 0.16 | 0.11 | 0.16 | 0.11 | 0.12 | 0.16 | 0.16 | 0.12 | 0.15 | 0.08 | 0.16 | 0.13 | 0.11 | 0.11 | 0.11 | 0.10 | 0.14 | 0.15 | 0.14 | 0.06 | 0.09 | 0.09 | 0.10 | 0.09 | 0.15 | 0.11 | 0.11 | 0.16 | 0.08 |

Note: For each evaluated country (rows), the columns show the average of the slacks amounts for each indicator—excluding from the average those subsamples in which the optimal slack solution is zero.

Reading Table 6 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, Austria identifies itself as benchmark peer in 75.6% 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 (7). This shows that the main diagonal identifies the number of times that the country under evaluation is effective by defining the utility frontier. Understandably, Denmark, the best country, is its own peer most of the times, 98.3%. On the contrary, still within the main diagonal, Latvia is the country that is deemed underperforming most frequently, as it is never efficient—even though it is not the most utility ineffective—followed by Croatia which is efficient just 1.9% of the times.

Focusing again on Austria, the first column reveals that although it performs relatively well—ranking 6th in $\overline{UE}^{LR}$—it is not frequently recognized as a technical peer. The last four rows summarize peer frequency statistics across columns, excluding self-assessments on the main diagonal. On average, Austria is selected as a benchmark in just 6.77% of the samples, placing it around the overall mean. In contrast, Denmark, the top performer, is chosen in 19.21% of subsamples, followed by Sweden (13.17%), Finland (10.71%) and Ireland (10.31%), all exceeding the 10% mark. Also, it is possible to determine which country is more relevant in a bilateral comparison taking their respective frequencies as reference peers. For example, considering the best and worst performing countries, Denmark and Romania, the former is peer to the latter in 27.7% of the samples. On the contrary, Romania is never a peer to Denmark.

This technical utility analysis helps pinpoint which countries serve as performance references from a technical perspective when calculating the radial component of technical effectiveness $\phi_k^*$ through program (7). However, when considering the strongly efficient frontier as reference to calculate $\overline{UTE}_k^S$, as we report in Tables 3 and it is relevant to consider the magnitude of the individual slacks, $s_{oi}^+$, for all $i = 1, \ldots, 32$, indicators. Indeed, the more pervasive is the presence of slacks when solving program (8), the greater the disparity between the utility technical effectiveness considering either the weakly or strongly efficient frontiers, i.e., $\overline{UTE}_k^W$ vs. $\overline{UTE}_k^S$.

Table 7 presents, for each country, the average value of the slacks for all 50,000 samples excluding the number of times when the optimal solution is zero. Recall that the values are normalized between zero and one, and therefore the slack measures the proportion in which each indicator can be improved within that range. For countries whose slacks are always zero, this value is reported in the corresponding cell. Considering Denmark as the best performing country that is normally technically efficient with $\phi_k^* = 1$, we observe in its corresponding row that there are no slacks in many indicators, while the maximum average slack, 0.29, corresponds to indicator 3.2.1 (Innovative SME collaborating with others—see Annex 1). On the contrary, Romania, always exhibits positive slacks—except for indicator 4.2.1 (Medium and high-tech product exports)—with a maximum value of 0.52 in indicator 1.2.1 (International scientific co-publications). Thanks to this information, each country can determine what individual dimensions are hampering their performance when attaining the performance level of their technical peers. Looking now at the descriptive statistics of the indicators, we see that the largest average slack across all countries is 3.2.1 (Innovative SMEs collaborating with others), with a value of 0.28, followed by 1.2.1 (International scientific co-publications) and 2.2.1 (R&D expenditure in the business sector) with 0.27. On the contrary, the indicator with the minimum average slacks is 4.2.1 (Medium and high-tech product exports) with 0.09. The general presence of slacks and their relatively high values confirm the relevance of decomposing utility effectiveness against the strongly efficient frontier that considers individual adjustments, rather than the weakly frontier that only considers radial increases.

### 4.3.3. *Benchmarks maximizing utility*

Once a country under evaluation increases its innovation indicators in the radial and individual slack amounts necessary to reach the utility frontier—thereby ensuring that it is not Pareto dominated—the final step to maximize its utility is to mirror the innovation system that, given the weights chosen by the DM, maximize the utility across all countries. As for the case of technical utility effectiveness, it is possible to determine the country(ies) that serve as reference when maximizing utility by looking at the optimal values of their associated multipliers. In this case, one is interested in the positive values of the multipliers in program (5), i.e. $\lambda_k^* > 0$, indicating that country $k$ is now involved in the determination of the MPA.

Reading now Table 8 below by columns, we report the number of times (in percentage) that a country maximizes the utility in the 50,000 samples. The table is organized following the categorization of countries proposed by the EIS (i.e., from 'Innovation leaders' to 'Emerging Innovators'). We see that Denmark is the country that maximizes the utility most frequently in all specifications, surpassing Sweden, Finland, and the Netherlands in all specifications, and accounting for a remarkable 76.2% of the times under the fixed proportions (Leontief) formulation. Therefore, Scandinavian and Northern European countries consistently emerge as role models. However, countries like Ireland—experiencing rapid economic growth—and smaller nations

**Table 8**
Benchmark peers maximizing utility (MPAs), $\lambda_{ki}^* > 0$ in program (5), EIS, 2024.

| | **Innovation leaders** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Utility | Denmark | Sweden | Finland | Netherl. | | | | | |
| Linear | 45.1 | 34.9 | 17.4 | 15.0 | | | | | |
| Cobb-Douglas | 62.3 | 46.0 | 24.8 | 26.9 | | | | | |
| Leontief | 76.2 | 55.0 | 38.5 | 37.2 | | | | | |
| | **Strong Innovators** | | | | | | | | |
| Utility | Belgium | Ireland | Luxemb. | Cyprus | Austria | Germany | Estonia | France | |
| Linear | 15.0 | 7.5 | 6.7 | 4.7 | 4.6 | 2.8 | 1.2 | 0.6 | |
| Cobb-Douglas | 27.6 | 12.9 | 12.2 | 9.1 | 13.4 | 7.8 | 4.5 | 2.3 | |
| Leontief | 41.6 | 31.6 | 35.4 | 33.6 | 47.3 | 35.2 | 30.0 | 12.9 | |
| | **Moderate Innovators** | | | | | | | | |
| Utility | Malta | Lithuania | Czechia | Italy | Spain | Greece | Portugal | Slovenia | Hungary |
| Linear | 0.5 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.0 | 0.0 | 0.0 |
| Cobb-Douglas | 1.3 | 0.7 | 0.6 | 0.5 | 0.3 | 0.2 | 0.1 | 0.0 | 0.0 |
| Leontief | 16.2 | 10.3 | 15.1 | 12.2 | 8.3 | 14.8 | 3.8 | 4.4 | 0.7 |
| | **Emerging Innovators** | | | | | | | | |
| Utility | Poland | Slovakia | Bulgaria | Croatia | Latvia | Romania | | | |
| Linear | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | | | |
| Cobb-Douglas | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | | | |
| Leontief | 1.4 | 0.9 | 0.7 | 0.1 | 0.0 | 0.0 | | | |

such as Cyprus, which are making relevant progress in innovation strategies, also frequently serve as peers. For underperforming EU countries, selecting a benchmark should consider the feasibility of replicating the innovation practices of leading nations, while recognizing that socioeconomic and institutional contexts may not be easily transferable.

We select Romania as the worst performing country to propose a typical strategy for innovation effectiveness improvements. Considering the case of the geometric mean represented by the Cobb-Douglas specification to depart from the traditional linear aggregator, this country is technically ineffective ($\overline{UTE}^{S}_{Romania} = 0.306$, Table 3) and could increase its utility on average by 226.8% (=(1/0.306−1)×100). This can be achieved by mirroring the innovation system of Denmark, which is its most frequent benchmark on the utility frontier: 27.7% of the times (Table 6). This technical improvement in terms of utility can be achieved by radially increasing all indicators by 79.6% ($\phi^{*}_{o} = 1.796$), as well as individual increments of individual indicators. As reported in Table 7, absolute increases in the value of several indicators—interpretable as percentages—are feasible on average, such as 0.52 in indicator 1.2.1 (International scientific co-publications), 0.51 in 2.1.1 (R&D expenditure in the public sector), and 0.48 in both 2.2.3 (Innovation expenditures per person employed in innovation-active enterprises), 3.1.1 (SMEs with product innovations), 3.2.2 (Public-private co-publications) and 4.1.1 (Employment in knowledge-intensive activities). The relevant magnitude of these individual adjustments shows the important efforts that Romania must make to improve its innovation system, as well as the relevance of taking into consideration the presence of slacks when evaluating utility technical effectiveness.

Once the technical frontier is reached, utility allocative effectiveness ($\overline{UAE}^{S}_{Romania} = 0.772$, Table 3) shows that an additional increase in utility to the tune of 29.5% (=(1/0.772−1)×100) is possible by balancing the values of indicators to match those of, once again, Denmark, which is the MPA maximizing utility 62.3% of the times (Table 8). Overall, by increasing and balancing its innovation indicators, Romania would be able to increase its utility in the eyes of the DM by 323.7% (=(1/0.236−1)×100) (i.e., $\overline{UE}_{Romania} = 0.236$, Table 3), an achievement that would require large investments and restructuring of the innovation system, for which this analysis provides a robust guidance.

From a standard BoD perspective, Romania's improvement strategy would focus on increasing all indicators to reach the frontier. However, the proposed utility-based framework provides an additional layer of insight: beyond increasing the level of indicators, Romania must also rebalance its innovation system to align with the preference structure embodied in the utility function. This highlights a key difference with the BoD approach, which focuses exclusively on quantity improvements, whereas our model also identifies misallocation across indicators as a source of ineffectiveness.

## 5. Conclusions and discussion

Composite indicators are predominantly used as policy benchmarking and communication instruments: they compress multidimensional information into a single message that can be compared across jurisdictions, tracked over time, and used to structure peer learning and policy debate [5,6]. Precisely because they serve political communication and monitoring purposes, their construction is inevitably contested—indicator selection, normalization, aggregation and weights embed value judgements and may be perceived as 'rigging' [7,9]. In this study we therefore do not pursue the ill-posed objective of identifying "scientifically correct" weights. Instead, we address a practical interpretability problem: when a country performs poorly under a CI, how much of the gap reflects objective shortfalls in indicator levels, and how much reflects subjective choices about substitutability and weights?

We answer this question by making the preference content of CIs explicit. Interpreting a CI as a DM's utility function over indicators [12,13], we define utility effectiveness as the ratio between the observed utility of an alternative and the maximum utility attainable within the observed feasible set (the Most Preferred Alternative, MPA). We work with the CES family because it nests the three canonical aggregators used in CI practice—linear (arithmetic mean), Cobb–Douglas (geometric mean) and Leontief (fixed proportions)—and thus spans the range from fully compensatory to non-compensatory evaluation [6,13]. Each specification is characterized by a specific degree of substitutability between indicators, with the linear specification being the most flexible by allowing perfect substitution and the fixed proportions penalizing the most any deviation from the weights chosen by the DM. The geometric mean specification also penalizes the dispersion between indicators, thereby providing an alternative to ad-hoc formulations that rescale indicators based on the dispersion in values observed among them, and then aggregates these penalized values using the usual arithmetic mean; i.e., the 'bottleneck' literature represented by Acs et al. [63] and Zofío et al. [58]. As the main implication of the study, we show that the choice of utility function by the DM is not neutral and can lead to relevant changes not only in the MPA, but in also the ranking of observations. Therefore, the choice of (un)penalizing function is critical in the evaluation process.

Our main methodological contribution is a decomposition of overall utility effectiveness into two mutually exclusive components with distinct institutional interpretations. First, utility-technical effectiveness (UTE) captures the objective, quantity-driven performance gap: it measures the distance to the best-practice frontier and coincides with (the inverse of) the DEA–BoD effectiveness score [9,14]. This component is independent of the particular utility specification, its substitutability parameter, and the weights, and therefore isolates the part of underperformance that is robust to the political contestation surrounding CI design. Second, utility-allocative effectiveness (UAE) captures the subjective, preference-driven gap: conditional on reaching the frontier, it measures the remaining utility loss due to an unbalanced mix of indicators relative to the DM's trade-offs embedded in the aggregator and weights.

This separation directly addresses the institutional logic of CI use. In policy discussions, frontier distance (low levels of indicators) and composition (a mismatch with the balance emphasized by the CI) are often conflated. Our decomposition disentangles them and clarifies the policy message. When UTE dominates, the implication is that poor performance reflects a genuine deficit in indicator quantities relative to best practice; debates about alternative weights or aggregators are second-order. When UAE is sizeable, rebalancing the internal composition of the system becomes central and the normative question of "which balance" should be prioritized legitimately comes to the fore. Making this dependence explicit improves transparency, and reduces the risk that disagreements over weights overshadow the substantive diagnosis of performance.

Beyond interpretation, the framework yields actionable benchmarking guidance consistent with how CIs are used for policy learning—see the discussion on Romania closing the previous section. A three-stage improvement logic emerges: (i) close the technical gap by expanding indicator levels to reach the frontier (DEA–BoD benchmarking); (ii) address non-radial shortfalls through slacks to identify which dimensions lag most [15]; and (iii) once on the frontier, reduce allocative losses by rebalancing indicator composition toward the MPA implied by the DM's utility structure. This staged interpretation avoids a common pitfall in CI discourse—treating any performance shortfall as a 'weighting problem' when it may primarily be a frontier-distance problem. This strategy is generally coherent even if the increments and changes required to become technically and allocative efficient might be contradictory, carrying unwanted adjustment costs. For this purpose, models directly projecting ineffective observations onto the MPA, identifying ineffectiveness as either technical or allocative, might be preferred [64–66].

A second practical contribution concerns feasibility in high-

dimensional CI settings. Standard DEA–BoD applications can suffer from the curse of dimensionality when the number of indicators is large relative to observations, weakening discrimination [67]. We therefore implement the approach with Ensemble-DEA [38], which preserves the full indicator set while improving robustness via repeated subsampling of countries and indicators. Relative to common remedies—pre-selecting indicators, collapsing them into higher-level pillars, or applying dimensionality-reduction techniques—this strategy maintains interpretability and avoids imposing additional ad hoc aggregation beyond the DM's stated preferences [57,58].

Empirically, considering the evaluation of innovation systems using data from the EIS, the linear arithmetic mean and the Cobb-Douglas geometric mean yield comparable results on average. Overall utility ineffectiveness is largely caused by the technical component, indicating that the main driver of underperformance is insufficient indicator levels rather than aggregation details. On the contrary, the fixed proportion Leontief, preventing any substitutability by definition, is very dissimilar, and identifies allocative ineffectiveness as the main cause of ineffective performance. This expected result explains why it is not adopted by international organizations, public institutions, or research bodies constructing CIs. Our illustration turns the 'weights debate' into an empirical question: it can matter in principle, but its practical impact can be assessed only once objective frontier distance is separated from preference-driven composition effects [9–11]. Nevertheless, despite the choice of aggregating function, we identify the same geographical divides in performance between European regions, with Scandinavian, Northern European and Benelux countries at the top of the innovation rankings and Southern and Eastern European countries exhibiting the worst performance.

Several extensions are natural. Future work may incorporate alternative technical effectiveness measures beyond the radial model (e.g., additive or directional measures) within the same decomposition logic [17,45,68]. The analysis of heterogeneous alternatives grouped by shared characteristics can be effectively addressed using meta-frontier approaches, which distinguish between intra- and inter-group effectiveness [25]. Extending the concept of utility effectiveness to a temporal framework—analogous to the Malmquist productivity index—and its decomposition into frontier change and effectiveness change would, in turn, enable the examination of performance dynamics over time [69]. From the perspective of the stakeholders, it is also possible to relax the single-DM assumption by considering heterogeneous preference structures and studying how allocative effectiveness changes across them, e.g. using cross-efficiency methods, Aparicio and Zofío [70]. Finally, while we illustrate the framework for innovation systems, the method applies broadly to other policy domains where CIs are used mainly for monitoring and communication rather than for identifying structural causal relationships, e.g., UNDP's Human Development Index.

In conclusion, the proposed utility effectiveness framework improves the interpretability of CI-based benchmarking in political settings by explicitly separating what is objective (frontier distance in quantities) from what is unavoidably normative (trade-offs embedded in aggregation and weights). Rather than attempting to eliminate the political–scientific tension inherent to CI construction, the method makes it measurable and transparent, allowing analysts and policymakers to communicate more clearly which part of the performance message is robust to aggregation choices and which part depends on them.

## CRediT authorship contribution statement

**José Luis Zofío:** Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Writing – original draft, Writing – review & editing. **Juan Aparicio:** Formal analysis, Methodology, Writing – review & editing. **Javier Barbero:** Data curation, Funding acquisition, Investigation, Software, Visualization, Writing – review & editing. **Jon Mikel Zabala-Iturriagagoitia:** Conceptualization, Resources, Validation, Writing – review & editing.

## Acknowledgments

The authors are grateful to the editor and to the reviewers for the comments provided in an earlier version of this paper. J. Aparicio and J. L. Zofío thank the grant PID2022-136383NB-I00 funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU. J.M. Zabala-Iturriagagoitia acknowledges financial support from the Basque Government Department of Education, Language Policy, and Culture (IT 1429-22). This research was also partially supported by the CIPROM/2024/34 grant, funded by the Conselleria de Educación, Cultura, Universidades y Empleo, Generalitat Valenciana. J. Barbero and J.L. Zofío acknowledge financial support from Grant PID2022-138212NA-I00 funded by MICIU/AEI/10.13039/501100011033 and by ERDF, EU.; from the Community of Madrid/ESF under program PHS-2024/PH-HUM-530 (DiTeCaM-CM); and from the Government of the Principality of Asturias and the European Regional Development Fund (IDE/2024/000712).

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.seps.2026.102528.

## Data availability

Data is publicly available at the EIS page: https://research-and-innovation.ec.europa.eu/knowledge-publications-tools-and-data/publications/all-publications/european-innovation-scoreboard-2024_en.

## References

[1] Saisana M, Neves A, Nurminen M, Starnoni E, Alberti V, Moura C. Indices and scoreboards in EU policymaking. European Commission: Joint Research Centre (JRC), Policy Brief; 2022. JRC131074.
[2] European Commission. European innovation scoreboard 2024. Luxemburg: Publications Office of the European Union; 2024. Available: https://research-and-innovation.ec.europa.eu/statistics/performance-indicators/european-innovation-scoreboard_en.
[3] Freudenberg M. Composite indicators of country performance: a critical assessment. Paris: OECD Publishing; 2003.
[4] Bandura R. Composite indicators and rankings: inventory 2011. New York: Office of Development Studies at the United Nations Development Programme (UNDP); 2011.
[5] Saltelli A. Composite indicators between analysis and advocacy. Soc Indic Res 2007;81(1):65–77.
[6] OECD. Handbook on constructing composite indicators: methodology and user guide. Paris: OECD Publishing; 2008.
[7] Grupp H, Schubert T. Review and new evidence on composite innovation indicators for evaluating national performance. Res Pol 2010;39(1):67–78.
[8] Cherchye L, Lovell CK, Moesen W, Van Puyenbroeck T. One market, one number? A composite indicator assessment of EU internal market dynamics. Eur Econ Rev 2007;51(3):749–79.
[9] Cherchye L, Moesen W, Rogge N, Van Puyenbroeck T. An introduction to 'benefit of the doubt' composite indicators. Soc Indic Res 2007;82:111–45.
[10] Greco S, Ishizaka A, Tasiou M, Torrisi G. On the methodological framework of composite indices: a review of the issues of weighting, aggregation, and robustness. Soc Indic Res 2019;141(1):61–94.
[11] Fusco E. Potential improvements approach in composite indicators construction: the multi-directional benefit of the doubt model. Soc Econ Plann Sci 2023;85:101447.
[12] Munda G, Nardo M. Constructing consistent composite indicators: the issue of weights. Tech. rept. EUR 21834 EN, European Commission 2005.
[13] Mas-Colell A, Whinston MD, Green JR. Microeconomic theory. Oxford: Oxford University Press; 1995.
[14] Cherchye L, Moesen W, Van Puyenbroeck T. Legitimately diverse, yet comparable: on synthesizing social inclusion performance in the EU. J Common Mark Stud 2004;42(5):919–55.
[15] Sahoo BK, Acharya D. Constructing macroeconomic performance index of Indian states using DEA. J Econ Stud 2012;39(1):63–83.
[16] Cherchye L, Moesen W, Rogge N, Van Puyenbroeck T. Constructing composite indicators with imprecise data: a proposal. Expert Syst Appl 2011;38(9):10940–9.
[17] Fusco E. Enhancing non-compensatory composite indicators: a directional proposal. Eur J Oper Res 2015;242:620–30.
[18] Zanella A, Camanho AS, Dias TG. Undesirable outputs and weighting schemes in composite indicators based on data envelopment analysis. Eur J Oper Res 2015;245:517–30.
[19] De Witte K, Rogge N. To publish or not to publish? On the aggregation and drivers of research performance. Scientometrics 2010;85(3):657–80.

[20] De Witte K, Rogge N. Accounting for exogenous influences in performance evaluations of teachers. Econ Educ Rev 2011;30(4):641–53.
[21] Vidoli F, Fusco E, Mazziotta C. Non-compensability in composite indicators: a robust directional frontier method. Soc Indic Res 2015;122:635–52.
[22] Fusco E, Vidoli F, Sahoo BK. Spatial heterogeneity in composite indicator: a methodological proposal. Omega 2018;77:1–14.
[23] Fusco E, Vidoli F, Rogge N. Spatial directional robust benefit of the doubt approach in presence of undesirable output: an application to Italian waste sector. Omega 2020;94:102053.
[24] Karagiannis R, Karagiannis G. Intra-and inter-group composite indicators using the BoD model. Soc Econ Plann Sci 2018;61:44–51.
[25] Walheer B. Disentangling heterogeneity gaps and pure performance differences in composite indexes over time: the case of the Europe 2020 strategy. Soc Indic Res 2019;143:25–45.
[26] Karagiannis G. On aggregate composite indicators. J Oper Res Soc 2017;68(7):741–6.
[27] Rogge N. Composite indicators as generalized benefit-of-the-doubt weighted averages. Eur J Oper Res 2018;267(1):381–92.
[28] Balk BM, Zofío JL. Aggregation rules with applications. Available at Social Science Research Network, SSRN 2025. https://doi.org/10.2139/ssrn.5096596.
[29] Van Puyenbroeck T, Rogge N. Comparing regional human development using global frontier difference indices. Soc Econ Plann Sci 2020;70:100663.
[30] Aparicio J, Kapelko M. Enhancing the measurement of composite indicators of corporate social performance. Soc Indic Res 2019;114(2):807–26.
[31] Oliveira R, Zanella A, Camanho AS. The assessment of corporate social responsibility: the construction of an industry ranking and identification of potential for improvement. Eur J Oper Res 2019;278(2):498–513.
[32] Aparicio J, Kapelko M, Monge JF. A well-defined composite indicator: an application to corporate social responsibility. J Optim Theor Appl 2020;186(1):299–323.
[33] Ferreira DC, Caldas P, Varela M, Marques RC. A geometric aggregation of performance indicators considering regulatory constraints: an application to the urban solid waste management. Expert Syst Appl 2023;218. Article 119540.
[34] Koronakos G, Kritikos M, Sotiros D. A common weights multiplicative aggregation approach for composite indicators: the case of global city competitiveness index. Expert Syst Appl 2024;242:122543.
[35] Kapelko M, Ortiz L, Aparicio J. Comparing groups of units through composite indicators in a non-convex approach: corporate social responsibility for the food and beverage manufacturing industry. Ann Oper Res 2024;351:1427–50.
[36] Vidoli F, Fusco E, Pignataro G, Guccio C. Multi-directional robust benefit of the doubt model: an application to the measurement of the quality of acute care services in OECD countries. Soc Econ Plann Sci 2024;93:101877.
[37] Walheer B. A sequential benefit-of-the-doubt composite indicator. Eur J Oper Res 2024;316(1):228–39.
[38] Aparicio J, Kapelko M, Monge JF, Zofío JL. Enhancing the benefit of the doubt model through 'Ensemble-DEA': achieving the sustainable development goals. Expert Syst Appl 2025;296(Part B). Article 129010.
[39] Gaaloul H, Khalfallah S. Application of the "Benefit-Of-the-Doubt" Approach for the Construction of a Digital Access Indicator: a Revaluation of the "Digital Access Index". Soc Indic Res 2014;118(1):45–56.
[40] González E, Cárcaba A, Ventura J. Weight constrained DEA measurement of the quality of life in Spanish municipalities in 2011. Soc Indic Res 2018;136:1157–82.
[41] Rogge N, Self R. Measuring regional social inclusion performances in the EU: looking for unity in diversity. J Eur Soc Pol 2019;29:325–44.
[42] Joro J, Korhonen PJ. Extension of data envelopment analysis with preference information: value efficiency. New York: Springer; 2015.
[43] Barbero J, Zabala-Iturriagagoitia JM, Zofío JL. Is more always the better? On the relevance of decreasing returns to scale on innovation. Technovation 2021;107:102314.
[44] Farrell MJ. The measurement of productivity efficiency. J Roy Stat Soc 1957;(120):253–90.
[45] Pastor JT, Aparicio J, Zofío JL. Benchmarking economic efficiency: technical and allocative fundamentals. Cham: Springer; 2022.
[46] Arrow KJ, Chenery HB, Minhas BS, Solow RM. Capital-labor substitution and economic efficiency. Rev Econ Stat 1961;43(3):225–50.
[47] Aldaz JM. Self-improvement of the inequality between arithmetic and geometric means. J Math Inequalities 2009;3(2):213–6.
[48] Mariani F, Ciommi M, Recchioni MC. A new class of composite indicators: the penalized power mean. Eur J Oper Res 2024;317:1015–35.
[49] UNDP. Human development report: the real wealth of nations. New York. Development Programme: United Nations; 2010.
[50] Koopmans T. An analysis of production as an efficient combination of activities. In: Koopmans T, editor. *Activity analysis of production and allocation*. Cowles commission for research in economics, monograph, 13. New York: Wiley; 1951.
[51] Charnes A, Cooper WW, Rhodes E. Measuring the efficiency of decision making units. Eur J Oper Res 1978;2(6):429–44.
[52] Lovell CAK, Pastor JT. Radial DEA models without inputs or without outputs. Eur J Oper Res 1999;118:46–51.
[53] Prieto AM, Zofío JL. Evaluating effectiveness in public provision of infrastructure and equipment: the case of Spanish municipalities. J Prod Anal 2001;15:41–58.
[54] Karagiannis G. DEA models without inputs or outputs: a tour de force. In: Parmeter C, Sickles RC, editors. Methodological contributions to the advancement of productivity and efficiency analysis. New York: Springer; 2020.
[55] Cooper WW, Ruiz JL, Sirvent I. Choices and uses of DEA weights. In: Cooper WW, Seiford LM, Zhu J, editors. Handbook on data envelopment analysis, 2011. New York: Springer Science+Business Media; 2011.
[56] Van Puyenbroeck T. On the output orientation of the benefit-of-the-doubt-model. Soc Indic Res 2018;139:415–31.
[57] Edquist C, Zabala-Iturriagagoitia JM, Barbero J, Zofío JL. On the meaning of innovation performance: is the synthetic indicator of the innovation union scoreboard flawed? Res Eval 2018;27(3):1–16.
[58] Zofío JL, Aparicio J, Barbero J, Zabala-Iturriagagoitia JM. The influence of bottlenecks on innovation systems performance: put the slowest climber first. Technol Forecast Soc Change 2023;193:122607.
[59] Cooper WW, Seiford LM, Tone K. Data envelopment analysis: a comprehensive text with models, applications, references and DEA-solver software. Princeton, NJ: Princeton University Press; 2007.
[60] Balk BM, Zofío JL. Rules for consistent aggregation: with application to cost efficiency decompositions. Forthcoming. Eur J Oper Res 2026.
[61] Simar L, Zelenyuk V. On testing equality of distributions of technical efficiency scores. Econom Rev 2006;25(4):497–522.
[62] Li Q. Nonparametric testing of closeness between two unknown distribution functions. Econom Rev 1996;15:261–74.
[63] Acs ZJ, Autio E, Szerb L. National systems of entrepreneurship: measurement issues and policy implications. Res Pol 2014;43:476–94.
[64] Zofío JL, Pastor JT, Aparicio J. The directional profit efficiency measure: on why profit inefficiency is either technical or allocative. J Prod Anal 2013;40:247–66.
[65] Kapelko M, Oude-Lansink A, Zofío JL. Endogenous dynamic inefficiency and optimal resource allocation: an application to the European dietetic food industry. Eur J Oper Res 2022;303(3):1444–57.
[66] Pastor JT, Zofío JL, Aparicio J, Alcaraz J. The flexible reverse approach for decomposing economic inefficiency: with an application to Taiwanese banks. Econ Modell 2024;139. Article 106804.
[67] Nataraja NR, Johnson AL. Guidelines for using variable selection techniques in data envelopment analysis. Eur J Oper Res 2011;215(3):662–9.
[68] Aparicio J, Pastor JT, Zofío JL. Can Farrell's allocative efficiency be generalized by the directional distance function approach? Eur J Oper Res 2017;257(1):345–51.
[69] Zabala-Iturriagagoitia JM, Aparicio J, Ortiz L, Carayannis EG, Grigoroudis E. The productivity of national innovation systems in Europe: catching up or falling behind? Technovation 2021;102:102215.
[70] Aparicio J, Zofío JL. Economic cross-efficiency. Omega 2021;100:102374.

**Ph.D. José Luis Zofío**: www.joselzofio.net, José Luis Zofío is a Professor of Economics at Universidad Autónoma de Madrid. He is also Visiting Professor at Erasmus University where he collaborates with several academics from the Rotterdam School of Management. His research interests are related to measurement theory in economics, in particular the use of index numbers for efficiency and productivity analysis, as well as spatial economics and trade theory. From an empirical perspective he undertakes multidisciplinary research, publishing numerous articles in top fields journals and book chapters related to environmental economics, transportation, innovation, and regional science. His first book, *Benchmarking Economic Efficiency*, was published by Springer in 2022. He has made several contributions to computational economics, with several packages for Matlab and Julia that focus on data envelopment analysis and regression methods.

**Ph.D. Juan Aparicio**, Juan Aparicio is a Professor at the Department of Statistics, Mathematics and Information Technology at the Miguel Hernandez University (Spain), and head of the Center of Operations Research. His research interests include efficiency and productivity analysis combined with machine learning and data science. He has published and co-edited several books on performance evaluation and benchmarking using Data Envelopment Analysis, as well as more than 100 scientific papers in the most relevant international journals in the field. He is also associate editor of Omega, the International Journal of Management Science.

**Ph.D. Javier Barbero:** http://www.javierbarbero.net, Javier Barbero holds a MSc. in Quantitative Economic Analysis from Universidad Autónoma de Madrid in 2012. He has visited several universities including the London School of Economic and Political Science (LSE) and the Université du Québec à Montréal (ESG-UQAM). Currently he is Assistant professor in the Department of Economics at Universidad Autónoma de Madrid. Previously, he was an Economic Analyst at the Directorate for Growth & Innovation in the European Commission Joint Research Centre (JRC), located in Seville (Spain), as a member of the Territorial Data Analysis and Modeling (TEDAM) team. His research interests are spatial economics, regional economic development, innovation, and efficiency and productivity analysis. He is also interested in computational economics and has published several articles in top journals developing code for efficiency and productivity measurement.

**Ph.D. Jon Mikel Zabala-Iturriagagoitia**: Jon Mikel Zabala-Iturriagagoitia is Associate Professor at the University of Deusto in San Sebastian (Spain). His research skills and interests include innovation policy, innovation management, and the use of indicators to inform policy decisions and evaluations related to innovation. His research has received numerous recognitions and awards, having had a strong policy impact. Due to the impact of his research, he has been appointed as a member of the Basque Academy of Sciences, Arts and Letters (Jakiunde), the Spanish foundation for innovation (COTEC), and the United Nations Economic Commission for Europe Task Force on Innovation Policy Principles.