Expert Systems With Applications 218 (2023) 119540

Contents lists available at ScienceDirect

Expert Systems With Applications

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

A geometric aggregation of performance indicators considering regulatory constraints: An application to the urban solid waste management

Diogo Cunha Ferreira [a,*,1], Paulo Caldas [b,c], Miguel Varela [b], Rui Cunha Marques [a,1]

[a] CERIS, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, 1, 1049-001, Lisbon, Portugal
[b] Business and Economic School, Instituto Superior de Gestão, Av. Mal. Craveiro Lopes 2A, 1700-284, Lisbon, Portugal
[c] CEG-IST, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, 1, 1049-001, Lisbon, Portugal

ARTICLE INFO

Dataset link: https://www.ersar.pt/pt/publicacoes/relatorio-anual-do-setor

Keywords:
Composite indicators
Multiplicative directional benefit of doubt
Imperfect knowledge of data
Regulation
Solid waste management

ABSTRACT

There are several ways of aggregating partial performance indicators into composites, each of them with advantages but also shortcomings and caveats. The objectivity demanded for policy-making and regulation of utilities leads many researchers to build their analyses over the so-called Benefit-of-Doubt (BoD), which, in turn, results from the well-known Data Envelopment Analysis (DEA). In a compensatory manner, these models construct piecewise linear frontiers containing the benchmarks, but they disregard increasing marginal products and the existence of non-concavity regions, thus underestimating efficiency. Multiplicative approaches have been proposed to solve these problems; one of these is the geometric (rather than linear) aggregation of variables. But still they fail to solve some problems like the correct treatment of undesirable variables, the existence of regulatory constraints, and the existence of imperfect knowledge of data. Therefore, this paper builds upon a geometric aggregation of performance indicators and proposes some strategies to solve the aforementioned shortcomings of the existing models. The new framework is exemplified and tested with the Portuguese urban solid waste management utilities.

1. Introduction

Achieving composite indicators translating the relative performance levels of a set of Decision Making Units (DMUs) has been an exercise carried out by many researchers in a while (Sala-Garrido et al., 2021; Silva et al., 2020). Interestingly, Greco et al. (2019) verified that, until January 2017, nearly six hundred papers have been published on this topic. Such performance levels or scores are useful for management design or policy development, as well as for utility performance (de Castro-Pardo et al., 2022; Yao et al., 2020) and regulation purposes (Henriques et al., 2020; Pereira, Camanho, Figueira, & Marques, 2021).

However, this process of constructing performance scores can be either subjective or objective (Fusco, 2015). On the one hand, evaluating the performance of DMUs may imply asking to a gathering of experts or decision-makers about their (subjective) judgments regarding a set of criteria that somehow describe performance. These criteria are typically made operational by the so-called key performance indicators (KPIs) (Matos et al., 2021). Using such judgments, we may apply some well-known multiple criteria decision tools and obtain scores translating the performance of DMUs (Greco et al., 2016; Zhou, Fan, & Zhou, 2010). Subjectivity might, then, be present in that analysis, hampering the achievement of concrete strategies to enhance those entities' performance (Moers, 2005). In contrast, objective frameworks typically require little or no human perceiving about the relationship between those criteria, turning strategies more robust and based only on empirical data, provided that there is a well-defined single objective function as an adequate representation of the decision problem (Buchanan et al., 1998). Not belittling the subjective frameworks, the main advantage of objectivity is that the performance of each DMU is not based on personal judgments that may bias the outcomes. That way, if we try to maximize the overall composite indicator by optimizing the criteria weights per DMU, then we get an objective-based composite indicator that cannot be larger by choosing another weighting scheme, so the managers of that DMU cannot complaint for being misjudged by subjectivity over weights (Cherchye et al., 2011). This differential weighting scheme finds support in the words of Knox Lovell et al. (1995): "*equality*

* Corresponding author.
E-mail addresses: diogo.cunha.ferreira@tecnico.ulisboa.pt (D.C. Ferreira), paulo.caldas@isg.pt (P. Caldas), miguel.varela@isg.pt (M. Varela), rui.marques@tecnico.ulisboa.pt (R.C. Marques).

[1] Diogo Cunha Ferreira and Rui Cunha Marques are grateful for the Portuguese Foundation for Science and Technology (FCT) support through funding UIDB/04625/2020 from the research unit CERIS.

https://doi.org/10.1016/j.eswa.2023.119540
Received 15 September 2022; Received in revised form 9 January 2023; Accepted 9 January 2023
Available online 21 January 2023

across components is unnecessarily restrictive, and equality across [DMUs] *and through time is undesirably restrictive. Both penalize a* [DMU] *for a successful pursuit of an objective, at the acknowledged expense of another conflicting objective. What is needed is a weighting scheme which allows weights to vary across objectives, over* [DMUs]".

Bottom line, deciding for a subjective or an objective-based model for a composite indicator building is strictly linked to the weighting schemes (whether plural or not, data-driven or expert judgments based) and the aggregating procedures (whether compensatory or not) (Greco et al., 2019). It does not mean that one cannot have an objectivity-based data-driven model for composite indicators assessment including some expert judgments as the regulators'.

Among the objective methods for performance assessment based on KPIs, the so-called Benefit of Doubt (BoD) stands out (Cherchye et al., 2007). BoD is a model derived from Data Envelopment Analysis (DEA). The idea underlying BoD is that one can aggregate a set of KPIs using an additive model and a linear program in which the weights of KPIs are the variables to optimize (Cherchye et al., 2011). These weights are often interpreted as the degree of importance of each criterion (Verbunt & Rogge, 2018). The result is a composite indicator condensing the information carried out by the KPIs.

BoD has been used extensively in many sectors for composite indicators estimation. Examples include: healthcare (Ferreira & Marques, 2021; Matos et al., 2021; Pereira, Camanho, Marques, & Figueira, 2021; Shwartz et al., 2016), waste, drinking water and wastewater management (D'Inverno et al., 2021; Mauricio-Iglesias et al., 2020), education (Agasisti et al., 2019; Coco et al., 2020; Sahoo et al., 2017), economics (Lafuente et al., 2021, 2020), banking (Gulati et al., 2020), community well-being (Bernini et al., 2013; Sarra & Nissi, 2013), human development (Blancard & Hoarau, 2013), active ageing (Amado et al., 2016), and quality of life (Rogge & Van Nijverseel, 2019).

Like DEA, the standard BoD is a linear programming model, constructing a piecewise linear frontier where the best performers are positioned (Van Puyenbroeck, 2018). Despite its wide acceptance, and as reported in the literature, the linear-based DEA approach has several weaknesses: (*a*) it does not account for increasing marginal products (Kao, 2017); (*b*) it may underestimate the efficiency of some non-dominated DMUs (Tiedemann et al., 2011); (*c*) it does not account for outputs competing for inputs (Banker & Maindiratta, 1986); (*d*) it has poor performance when the production technology exhibits non-concavity regions (Emrouznejad & Cabanda, 2010); and (*e*) it suffers from the problem of zero-multiplier (Tofallis, 2014), which is to say that some variables may not be accounted for in the performance score computation. So, there is a compensability problem in these models. By extension, the linear-based BoD approach suffers from (some of) these shortcomings. Besides, the additive model (that rules BoD) is not admissible when KPIs are not mutually preferentially independent (Debreu, 1960; Keeney & Raiffa, 1993; Krantz et al., 1971). It may constitute another problem for the standard linear BoD model in many empirical applications as that condition may not be always verified in practice, for instance, because of multicollinearity (Fusco, 2015).

That way, a solution proposed by several authors is the so-called multiplicative DEA that geometrically aggregates variables and constructs a piecewise log-linear frontier (Liu et al., 2020; Mehdiloozad et al., 2014). This alternative is more flexible than the standard DEA regarding the marginal products, the frontier shape (which does not have to be strictly convex), the rates of substitution, the zero-multiplier (compensability) problem, and the underestimation of efficiency as all non-dominated DMUs are classified as efficient (Ferreira, Marques, & Nunes, 2021). Additionally, Zhou, Ang, and Zhou (2010) claim that this geometric aggregation presents better theoretical properties than other aggregation schemes, including scale-invariance and lower degree of compensation between variables, leading to a minimum loss of information compared to other aggregating methods.

Interestingly, there are several applications of this multiplicative framework to construct composite indicators for performance evaluation (Blancas et al., 2013; Dominguez-Gil et al., 2021; Emrouznejad & Cabanda, 2010; Giambona & Vassallo, 2014; Rogge, 2018a; Tofallis, 2014; Van Puyenbroeck & Rogge, 2017; Verbunt & Rogge, 2018). Van Puyenbroeck and Rogge (2017) and Verbunt and Rogge (2018) classify these applications in two main groups: direct and indirect approaches. In short, these streams differ in both weighting and aggregation: the former combines them into a single multiplicative model, whereas the latter uses two steps, the first of which based on the linear BoD as instrument for assessing weights, followed by the composite indicator construction using a weighted geometric mean with the previously estimated weights.

Notwithstanding, the found studies fail in solving one (or more) of the following problems: (*f*) they disregard the existence of some undesirable or bad KPIs, that unequivocally exist in many empirical applications;[2] (*g*) when they do, the treatment follows the well-known weak disposability or an arguable rescalling of variables, already reported to be incorrect for modeling the production process (Podinovski & Kuosmanen, 2011); (*h*) the studies do not consider minimum and maximum levels often imposed by regulators (Marques et al., 2018), or any other stakeholder (Ferreira, Nunes, & Marques, 2021; Simões et al., 2010), such that DMUs falling below that minimum or above that maximum should not be considered best practices, which could occur with any BoD approach; (*i*) additionally, they also fail to recognize the existence of KPIs whose values should fall within certain ranges, also defined by regulators; and (*j*) those studies do not account for the problem of imperfect knowledge of data implying that the performance estimates are no longer deterministic, but stochastic instead (Ferreira, Marques, & Pedro, 2018). The main purpose of the current study is to provide a tool capable of solving the identified problems (*f*) to (*j*), through a multiplicative BoD approach, which, in turn, should be capable of surpassing the shortcomings (*a*) to (*e*). Sections 3 and 4 detail the proposed framework and present some illustrative examples for better understanding.

The case study of this paper focus on urban solid waste (SW) collection. Measuring the performance of operators in this sector has been repeatedly reported as essential in the context of a circular economy (Delgado-Antequera et al., 2021). Although the increase of SW production is linked to the economic development, such massive growth watched in the past few years pressures Governments for a better management of waste (Llanquileo-Melgarejo & Molinos-Senante, 2021). Indeed, the produced SW has meaningful consequences over green house emissions (Maria et al., 2020), public health (Ferreira, Graziele, et al., 2021; Mora et al., 2014), and the sustainability of natural resources (Yadav & Karmakar, 2020). Indeed, lockdowns resulting from the recent Covid-19 pandemics might have induced the production and disposal of urban SW (Cai et al., 2021). Thus, it is paramount to evaluate the performance of entities in collecting and disposing SW. Using a multiplicative BoD-like aggregating framework emerges as a new tool for SW regulators, as data tend to be gathered using KPIs. As detailed in the next section, there is only a handful of papers constructing composite indicators in this sector (e.g., Fusco et al. (2020), Mergoni et al. (2021), Olay-Romero et al. (2020), Ríos and Picazo-Tadeo (2021)). Those are mostly based on the standard BoD, thus suffering from the aforementioned shortcomings. Therefore, this research seeks to overcome the latter, applying the newly designed framework to evaluate the performance of SW collection companies in Portugal. Section 5 presents the case study and Section 6 the results and their discussion. The concluding remarks are left for the last section of this paper (see Section 7).

Based on Tsolas et al. (2020) and Charles et al. (2019), we frame our approach in view of Design Science Research Methodology (DSRM) (Peffers et al., 2008), a sequence of activities to be carried out so as

[2] In some cases, the authors tend to transform undesirable into desirable KPIs using translation and scale transformations. However, most of the analyzed multiplicative models are not either scale or translation invariant.

Table 1.1
Design Science Research Methodology (DSRM) applied to the current study.
Source: Adapted from the works of Charles et al. (2019), Tsolas et al. (2020), and Peffers et al. (2008).

DSRM activities	Activity description	Knowledge
Problem identification and motivation	Need to address the problem of regulatory constraints, undesirable KPIs, and imperfect knowledge of data when constructing composite indicators with BoD	Literature review. Understanding of shortcomings of existing BoD models for dealing with the identified problems. Understanding the potentiality of multiplicative BoD.
Define the objectives of a solution	Investigate whether it is possible to integrate regulatory constraints, undesirable KPIs, and imperfect knowledge of data in a multiplicative BoD model.	Literature review. Knowledge of existing tools: BoD, linear and multiplicative.
Design and development	Design of an integrated approach. This approach should be capable of returning targets within regulatory-defined intervals, while ensuring that undesirable KPIs are not incorrectly modeled and the imperfect knowledge of data is considered. It should be possible to categorize the DMUs according to their performance: benchmarks, potential benchmarks, and open to improvement.	
Demonstration	Case study demonstration: urban solid waste collection and management utilities. The proposed model is used to evaluate the performance of Portuguese urban solid waste entities, to classify them based on predefined efficiency classes, and to conduct some statistical inference over their performance.	Applying the proposed model to a real-world case study.
Evaluation	Comparative analysis.	Understanding of current solution and its advantages, disadvantages, and caveats.

to get into the artefact, which is the development of the proposed multiplicative BoD model with regulatory constraints, proper treatment of undesirable KPIs, and imperfect knowledge of data accounting. Table 1.1 presents the relevant activities and knowledge, which correspond to the essential materials required to conduct those activities. This study follows the DSRM activities to design and evaluate the suggested model.

2. Literature review

2.1. Composite indicators using the BoD approach

Since (Cherchye et al., 2007) have proposed the BoD approach as an output-oriented DEA with unitary inputs, several papers have emerged. Table 2.1 identifies and briefly describes some selected developments on BoD linear approach. We can identify six distinct streams:

1. Conservatism in weights estimation for KPIs aggregation (Zhou et al., 2007), through the construction of the worst scenario for DMUs;
2. Integration of BoD and multicriteria decision analysis (Guardiola & Picazo-Tadeo, 2014) to find common weights for all DMUs;
3. Inclusion of a directional nature into BoD (Fusco, 2015; Fusco et al., 2018; Zanella et al., 2015) such that the KPIs do not expand all at the same rate to project the DMU into the frontier;
4. Inclusion of non-discretionary variables into BoD (Fusco et al., 2018) that help explaining performance (although not under the control of managers);
5. Reducing the impact of outliers and extreme values in performance estimation (Fusco et al., 2020, 2018), through the merging of BoD and order-*m*;
6. Inclusion and treatment of undesirable (or reverse) KPIs in BoD (Färe et al., 2019; Zanella et al., 2015).

Neither of these studies have dealt with the problem of fixing regulatory-imposed restrictions like minimum or maximum thresholds, below or above which one cannot be considered a benchmark. Although some have considered the problem of including undesirable KPIs, the most common treatment follows the weak disposability or an arguable rescaling of variable, which is an incorrect modeling according to Podinovski and Kuosmanen (2011).

2.2. Composite indicators using the multiplicative BoD approach

Table 2.2 presents an overview of the works on the multiplicative versions of BoD. Building on the disadvantages of the linear BoD, two main streams have emerged on the multiplicative BoD. They differ essentially in weighting and aggregation. The first stream considers a simple geometric mean-like aggregation, which implies that the logarithm of KPIs needs to be considered for linearization purposes. Weights are simple artifacts of the model. The second stream starts by using the linear BoD to optimize weights, which will be used in a weighted geometric mean of KPIs. For the authors of this paper, the rationale of using weights endogenously computed via a linear model as inputs of a geometric model for aggregation is hard to follow, especially when such an aggregation can be directly performed in a single model. Besides, such a procedure is likely to bias the weights estimation because of the compensatory nature of the linear BoD.

3. A log-BoD approach to construct a composite indicator when all data are deterministic and perfectly known

Let us consider a comprehensive set of s strictly positive KPIs, $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_r, \ldots, \boldsymbol{y}_s)^\top$, evaluated for n DMUs, *i.e.*, $\boldsymbol{y}_r = (y_r^1, \ldots, y_r^j, \ldots, y_r^n)^\top$ for any $r = 1, \ldots, s$. Generically, the $s \times n$ matrix containing all data is represented as follows:

$$\boldsymbol{Y} = \begin{pmatrix} y_1^1 & \cdots & y_r^1 & \cdots & y_s^1 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ y_1^j & \cdots & y_r^j & \cdots & y_s^j \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ y_1^n & \cdots & y_r^n & \cdots & y_s^n \end{pmatrix}.$$

Note that $y_r^j > 0$ for any $r = 1, \ldots, s$ and $j = 1, \ldots, n$. In the present study, each DMU represents a particular Portuguese urban solid waste collection entity.

Table 2.1
Some selected developments on BoD linear approach.

Author(s) and study	Brief description of the study
Cherchye et al. (2007), An introduction to benefit of the doubt composite indicators	Based on Melyn and Moesen (1991), this seminal work presents the linear BoD that, bottom line, is an output-oriented DEA with unitary inputs. A set of KPIs is aggregated using a weighted arithmetic mean, such that weights are DMU-specific, and the overall or composite indicator should be smaller than or equal to one. The objective function of the associated linear program is to maximize the weighted average of the DMU whose performance is being assessed. That way, a DMU would maximize the weights that would benefit its overall performance, i.e., the weights of criteria in which the DMU performs well. In opposition, the criteria in which the DMU performs poorly would have weights close (or even equal) to zero. This is, clearly, a compensatory benevolent model.
Zhou et al. (2007), A mathematical programming approach to constructing composite indicators	In opposition to Cherchye et al. (2007), Zhou et al. searched for a conservative model, in which the "worst" set of weights for each entity are used to aggregate the KPIs into a performance score. The model resembles the input-oriented DEA model with unitary outputs. Therefore, the proposed model measures how far each DMU is from the worst practices frontier. Finally, the authors propose an index that results from the linear scaling in the min–max range, i.e., adding the composite indicator of Cherchye et al. and the one obtained using the proposed model, weighted by the adjusting parameters L and 1 – L (L between 0 and 1), respectively. Although the approach is interesting, the choice of parameter L is rather arbitrary.
Guardiola and Picazo-Tadeo (2014), Building weighted-domain composite indices of life satisfaction with data envelopment analysis	Guardiola and Picazo-Tadeo criticize the approach of Cherchye et al. (2007) because of the flexibility in determining weights, which is an advantage for some. The reasons identified for such criticism are three: first, according to the authors, different sets of weights prevent comparing individuals on a common basis; second, there is a potential lack of discriminatory power of the BoD, especially whenever there are too many KPIs compared to the number of DMUs; and third, there is a problem of compensation resulting from null weights. The authors propose the combination of BoD and multiple criteria decision-making tools to obtain a common set of weightings via the so-called Global-Efficiency approach of Despotis (2002) and Bernini et al. (2012). Such an approach depends on a parameter of the objective function that ranges from zero to one: if it is zero, the model corresponds to maximizing the deviation between the BoD scores and the score computed with a common scheme of weights, being thus the most penalizing optimum. In opposition, if that parameter is one, the model minimizes the average deviation across DMUs between their BoD scores computed with the most self-favorable weighting scheme and the scores with common weightings. However, the choice of such a parameter is an arbitrary exercise. Besides, sets of common weights decrease the interest of a BoD approach as the flexibility is the reason for its wide acceptance (Decancq & Lugo, 2013; Greco et al., 2019).
Zanella et al. (2015), Undesirable outputs and weighting schemes in composite indicators based on data envelopment analysis	This paper explores the construction of composite indicators using a directional BoD approach, like the one of Fusco (2015), but accounts for undesirable KPIs. The authors consider two distinct procedures for dealing with this kind of variable: direct and indirect. The direct approach considers the directional BoD model while imposing weak disposability for undesirable KPIs and strong disposability for the remaining dimensions (later on, the authors relax the weak disposability, treating the undesirable KPIs as inputs in a directional DEA model). The indirect approach, built on Seiford and Zhu (2002), uses a very large quantity such that the model penalizes less the DMUs with more significant differences between that quantity and the undesirable KPIs.
Fusco (2015), Enhancing non-compensatory composite indicators: A directional proposal	Fusco extends the original BoD model, proposing a directional version of it, in which one uses a vector defining the path to project each DMU in the frontier. Such an extension results from Chambers et al. (1998) and Bogetoft and Otto (2011). The problem here lies in defining the most suitable vector for that path. Although there are several alternatives in the literature (see Bogetoft & Otto, 2011), Fusco considers them as requiring experts' judgments. Thus, to overcome such an issue, which is regarded as a drawback, the author proposes using principal component analysis to discover the endogenous preference structure among KPIs. According to the author, this alternative allows deriving both the preference structure and the direction from the data avoiding subjective judgments.
Rogge (2018b), On aggregating Benefit of the Doubt composite indicators	Building his work on Färe and Zelenyuk (2003) and Färe and Karagiannis (2014), Rogge aggregates individual composite indicators into groups of composites, as, in some practical situations, the performance analysis of a group of DMUs could be more interesting than the individual examination. The author essentially applies a BoD to aggregate composite indicators themselves (a similar exercise was conducted by Matos et al., 2021).
Fusco et al. (2018), Spatial heterogeneity in composite indicator: A methodological proposal	In this paper, Fusco and colleagues try to account for non-discretionary dimensions directly in the BoD model to construct a composite indicator adjusted for those variables. Reasonably, the authors consider that capturing the effect of exogenous variables is mandatory for a well-performed construction of composite indicators. This fact was widely discussed in the literature before, however with respect to the standard DEA models. The position of DMUs concerning the frontier tends to be conditioned by these dimensions. In short, Fusco et al. integrate the BoD with the well-known order-*m* method of Cazals et al. (2002). Note, however, that such an adjustment demands for a kernel function and a bandwidth, and their choice can be debatable as results tend to be sensitive (Ferreira, Marques, & Nunes, 2018).
Färe et al. (2019), A benefit-of-the-doubt model with reverse indicators	This work proposes a new version of BoD to account for undesirable KPIs (or reverse indicators in the terminology of the paper). In short, after criticizing the transformation of the variables used by some and the directional BoD alternatives, both to deal with the problem of undesirable KPIs, the authors present a linear programming model that basically treats these variables as inputs and desirable KPIs as outputs in a basic DEA model.
Fusco et al. (2020), Spatial directional robust Benefit of the Doubt approach in presence of undesirable output: An application to Italian waste sector	The authors extend the work of Fusco et al. (2018) to account simultaneously for non-discretionary variables that describe the exogenous operational environment and the existence of undesirable KPIs. In that sense, Fusco et al. integrate the directional BoD (that allows to account for undesirable KPIs) with a Monte-Carlo cycle like the one of order-*m* (Daraio & Simar, 2014). In each iteration, the DMU is compared to *m* other DMUs randomly selected but operating in the same environment of the former. That way, the results become adjusted by the operational environment that may somehow condition the performance of DMUs. Later, D'Inverno and De Witte (2020) conducted a similar exercise, this time focusing on service-level provision in municipalities.

As one may have noticed from the literature review, most of the aggregations carried out with BoD rely on the multiplier model, trying to optimize the weights associated with each criterion. However, those models do not provide insights about the targets per KPI and DMU. Besides, in regulatory term, instead of fixing lower and upper levels for weights, it is easier to define empirically-based boundaries for KPIs to claim that a DMU is a good or a poor performer in each dimension. So, instead of working with aggregation along all criteria, we construct a (multiplicative) model that aggregates the data of all DMUs and a single KPI. It leads us to the definition of *target*. Indeed, we can define the performance of a DMU $k \in \{1, \ldots, n\}$ as the relationship between the observed values and the corresponding optimal values in all KPIs. Let us impose that the optimal value associated with the indicator $\boldsymbol{y}_r$ of the DMU k results from the combination of the observed values for its possible benchmarks: $(y_r^k)^*$ denotes this optimum for $r = 1, \ldots, s$. Let $\boldsymbol{\mu}^k = (\mu_1^k, \ldots, \mu_j^k, \ldots, \mu_n^k)$ be a set of n non-negative coefficients weighting the contribution of each benchmark to the optimum $(y_r^k)^*$, provided that $\sum_j \mu_j^k = 1$. If the DMU k is a benchmark itself, we have $\mu_k^k = 1$

Table 2.2
Some selected developments on BoD multiplicative approach.

Author(s) and study	Brief description of the study
Emrouznejad and Cabanda (2010), An aggregate measure of financial ratios using a multiplicative DEA model	Based on the work of Banker and Maindiratta (1986), Emrouznejad and Cabanda propose a multiplicative model aiming to maximize a score (composite indicator) provided that the targets associated with KPIs are aggregated a weighted geometric mean with weights to be optimized. If such an aggregation would follow weighted arithmetic mean, the model would be a DEA with no inputs. Although the initial model is not linear, a simple transformation of variables using logarithms erases that problem.
Zhou, Ang, and Zhou (2010), Weighting and aggregation in composite indicator construction: A multiplicative optimization approach	In their study, Zhou et al. propose a weighted geometric aggregation of KPIs, such that weights are to be optimized using a linear program. The aggregation is like the one of Emrouznejad and Cabanda (2010), with the difference that the right-hand side of the main constraint is the Napier number. Just as Zhou et al. (2007), the authors propose two distinct models, one that searches for the distance to the best frontier and another for the distance to the worst frontier. Then, it is proposed to aggregate both best and worst composite indicators using linear scaling in the min–max range and a parameter L (from 0 to 1), whose choice is rather arbitrary, as previously stated.
Blancas et al. (2013), Constructing a composite indicator with multiplicative aggregation under the objective of ranking alternatives	In this study, Blancas et al. propose a common-weight approach that tries to avoid ties between alternatives, which is, according to the authors, a desirable property, especially when one wants to construct rankings. The model is based on a mixed-integer linear program that searches for the minimum distance from an "ideal" composite indicator and a composite obtained as a model's variable. The latter results from a geometric aggregation as in Emrouznejad and Cabanda (2010) and Zhou, Ang, and Zhou (2010). The main criticism that can be made to this study is the fact that composite indicators result from a fixed, although endogenously computed, set of weights, while the flexibility of weights is the 'beauty' of this kind of models.
Tofallis (2014), On constructing a composite indicator with multiplicative aggregation and the avoidance of zero weights in DEA	The author criticizes Blancas et al. (2013) and Giambona and Vassallo (2014) for a normalization step that removes proportionality and destroys the desirable ratio scale property. Then, the author proposes a score resulting from a Cobb–Douglas function with a scaling factor. By applying logarithms to data and using the transformed data into a linear programming model maximizing the sum of slacks, the author concludes that the resulting model avoids the zero-weight problem after obtaining its dual version.
Van Puyenbroeck and Rogge (2017), Geometric mean quantity index numbers with benefit-of-the-doubt weights	The model proposed by Van Puyenbroeck and Rogge is an indirect multiplicative BoD based on two steps. First, the authors use a linear BoD to estimate the weights of KPIs; and second, they use these weights in a weighted geometric average of the KPIs divided by an arbitrarily defined baseline. According to these authors, the indirect multiplicative BoD model avoids some of the disadvantages of the direct models of Zhou, Ang, and Zhou (2010), Giambona and Vassallo (2014), and Tofallis (2014), namely the lack of commensurability and the presence of scaling factors. However, since weights result from the linear BoD, they might be biased by the total compensation of sufficiently high values on some KPIs.
Verbunt and Rogge (2018), Geometric composite indicators with compromise Benefit-of-the-Doubt weights	The paper builds on Van Puyenbroeck and Rogge (2017)'s indirect multiplicative BoD to construct a geometric composite indicator and estimate the importance of each KPI. The authors combine optimistic and pessimistic BoD-based weighting to explain differences in policy performances under the different weighting schemes following a multiplicative Bortkiewicz decomposition.
Rogge (2018a), Composite indicators as generalized benefit-of-the-doubt weighted averages	This paper also builds on the two steps (Van Puyenbroeck & Rogge, 2017) to construct a composite indicator. In the first step, the author uses the linear BoD to estimate the weights for each DMU and all KPIs. In the second stage, a generalized mean is used with the weights previously computed to evaluate the impact of the Hölder order in the composite indicator construction. As that order tends to $-\infty$, the composite results from the KPI or KPIs in which the DMU has poor performance. In opposition, for larger orders, the model penalizes less the poor performance in some dimensions and, therefore, is less conservative as only the KPIs with good performance are considered.

and $\mu_j^k = 0$ for any $j \neq k$. In this case, $(y_r^k)^* = y_r^k$ for any $r = 1, \dots, s$. Otherwise, should $(y_r^k)^* \neq y_r^k$ for some $r = 1, \dots, s$, then the DMU k cannot be a benchmark.

In practice, we can use the $\mho$-Hölder average of the n observations associated with $\boldsymbol{y}_r$ and weights $\boldsymbol{\mu}^k$ to define the optimum $(y_r^k)^*$. Let us use the angular brackets to represent this average:

$$(y_r^k)^* = \langle \boldsymbol{\mu}^k \mid \boldsymbol{y}_r \rangle_{\mho} \overset{\text{def}}{=} \left\langle \boldsymbol{\mu}^k \mid (\boldsymbol{y}_r)^{\mho} \right\rangle^{\frac{1}{\mho}} = \left(\sum_j \mu_j^k (y_r^j)^{\mho} \right)^{\frac{1}{\mho}}, \ r = 1, \dots, s; \ \mho \in \mathbb{Z}_0. \tag{3.1}$$

In a recent research, Rogge (2018a) used a family of functions to aggregate normalized KPIs and, as such, construct a composite indicator. Such a family is ruled by the parameter $\mho$ of the Hölder average, such that $\mho$ ranges from $-\infty$ to $+\infty$. In short, if $\mho \to -\infty$, the resulting composite would be equal to the worst performance of the DMU in any KPI. In opposition, for $\mho \to +\infty$, the resulting composite would be the best performance of that DMU in any KPI. It does become obvious that larger values of parameter $\mho$ lead to higher compensation levels as the weaker KPI is the one with the smaller influence. Likewise, smaller values of $\mho$ tend to extremely penalize DMUs for their poor performance in some KPIs, ignoring good performance in other dimensions. Although this is linked to the nature of the decision problem and perhaps the decision-maker attitude, we believe that a neutral position is better, so $\mho = 0$.

Using the data in $\boldsymbol{Y}$ we can construct a frontier where benchmarks are located and against which the low performance DMUs are compared. Since $\mho = 0$ (zero-order Hölder average), we may construct a piecewise log-linear frontier, thus we coin the model as log-BoD, which is a multiplicative model based on a geometric aggregation (*vide infra*). Let us use the natural logarithm in any algebraic operations: $\log(a) = b \Longleftrightarrow a = \exp(b)$. It leads us to:

$$(y_r^k)^* = \langle \boldsymbol{\mu}^k \mid \boldsymbol{y}_r \rangle_0 \overset{\text{def}}{=} \prod_j (y_r^j)^{\mu_j^k} \Longleftrightarrow \log(y_r^k)^* \overset{\text{def}}{=} \sum_j \mu_j^k \log y_r^j = \langle \boldsymbol{\mu}^k \mid \log \boldsymbol{y}_r \rangle_1, \ r = 1, \dots, s, \tag{3.2}$$

provided that $\sum_j \mu_j^k = 1$ (Bullen, 2003, pp. 175–265). Note that this last condition is mandatory for a convex combination (Rockafellar, 1970, pp. 11–12), otherwise the weighted geometric mean in Eq. (3.2) would not be well-defined, especially when the variables in $\boldsymbol{Y}$ are limited empirically. For instance, suppose that a KPI is expressed as a percentage; thus, we know that its target cannot be beyond 100%. If the normalization condition does not hold, the target could be larger than this limit, which is not possible in practice. Indeed, building upon (Hardy et al., 1952), we can show that

$$\lim_{\mho \to -\infty} \langle \boldsymbol{\mu}^k \mid \boldsymbol{y}_r \rangle_{\mho} = \min_j y_r^j \leqslant \prod_j (y_r^j)^{\mu_j^k} \leqslant \max_j y_r^j = \lim_{\mho \to +\infty} \langle \boldsymbol{\mu}^k \mid \boldsymbol{y}_r \rangle_{\mho}$$

for any $r = 1, \dots, s$ if and only if $\sum_j \mu_j^k = 1$ (Du et al., 2014). Putting it differently, should $\max_j y_r^j \leqslant 100\%$ and $\min_j y_r^j \geqslant 0\%$, as happens to percentages, then the target of the KPI obviously ranges from 0 to 100%, thus being feasible in practice.

Not all indicators influence positively the DMUs' performance. KPIs can indeed be either *desirable* (or *good*), *neutral*, or *undesirable* (or *bad*). The higher the values of desirable or good KPIs, the better the performance. In opposition, the higher the undesirable or bad KPIs, the lower the performance. Meanwhile, neutral KPIs are neither desirable nor undesirable but are part of the production process and should be contained between pre-specified ranges, $\underline{\delta}_r$ to $\overline{\delta}_r$; DMUs whose observations fall outside these limits have a poor performance. Let s^D, s^N, and s^U be the number of desirable, neutral, and undesirable KPIs, respectively. As such, we have $s = s^D + s^N + s^U$. Likewise, let $\boldsymbol{y}_r^+$, $\boldsymbol{y}_r^0$, and $\boldsymbol{y}_r^-$ denote them, for any $r = 1, \ldots, s^D$ (desirable), $r = s^D + 1, \ldots, s^D + s^N$ (neutral), and $r = s^D + s^N + 1 \ldots, s$ (undesirable KPIs), respectively.

In some cases, regulators (or other decision-making entities) might fix thresholds as the minimum or maximum acceptable levels for desirable or undesirable KPIs, respectively. Let τ_r^+ be the threshold defining the minimum level for a DMU to be considered a benchmark for $r = 1, \ldots, s^D$ (desirable KPIs). Likewise, τ_r^- for $r = s^D + s^N + 1, \ldots, s$ is the maximum level for undesirable KPIs. If no thresholds have to be imposed, we may replace them as $\tau_r^+ = 0$ and $\tau_r^- = +\infty$ in what follows.

It should now be evident that the following relationships must hold:

(a) $(y_r^{k+})^* \geqslant \max\{y_r^{k+}, \tau_r^+\}$ for any $r = 1, \ldots, s^D$, *i.e.*, the optimal values of desirable KPIs should be at least equal to their observations and the regulator-defined thresholds;
(b) $\underline{\delta}_r \leqslant (y_r^{k0})^* \leqslant \overline{\delta}_r$ for any $r = s^D + 1, \ldots, s^D + s^N$, where $\underline{\delta}_r$ and $\overline{\delta}_r$ denote respectively the bottom and the top levels defined by stakeholders for neutral KPIs; and
(c) $(y_r^{k-})^* \leqslant \min\{y_r^{k-}, \tau_r^-\}$ for any $r = s^D + s^N + 1, \ldots, s$, *i.e.*, the optimal values of undesirable indicators should be at the most equal to their observations and the regulator-defined thresholds.

Considering the above three relationships, we have:

(a) **Desirable KPIs:**

$$\begin{aligned}&\langle \boldsymbol{\mu}^k \mid \boldsymbol{y}_r^+\rangle_0 \geqslant \max\{y_r^{k+}, \tau_r^+\}\\ &\quad\Longleftrightarrow \langle \boldsymbol{\mu}^k \mid \log \boldsymbol{y}_r^+\rangle_1 \geqslant \max\{\log y_r^{k+},\ \log \tau_r^+\} + Q_r^{k+} \Longleftrightarrow\\ &\quad\Longleftrightarrow \left\langle \boldsymbol{\mu}^k \;\middle|\; \log \frac{1}{\boldsymbol{y}_r^+} \right\rangle_1 + Q_r^{k+} \leqslant \min\left\{\log \frac{1}{y_r^{k+}},\ \log \frac{1}{\tau_r^+}\right\},\\ &\quad r = 1, \ldots, s^D,\end{aligned} \tag{3.3}$$

where Q_r^{k+} is a non-negative unknown quantity, and the inequality results from the convexity and the strong disposability of desirable criteria;

(b) **Neutral KPIs:**

$$\begin{aligned}&\langle \boldsymbol{\mu}^k \mid \boldsymbol{y}_r^0\rangle_0 \geqslant \underline{\delta}_r \ \wedge\ \langle \boldsymbol{\mu}^k \mid \boldsymbol{y}_r^0\rangle_0\\ &\quad\leqslant \overline{\delta}_r \Longleftrightarrow \langle \boldsymbol{\mu}^k \mid \log \boldsymbol{y}_r^0\rangle_1 \geqslant \log \underline{\delta}_r \ \wedge\ \langle \boldsymbol{\mu}^k \mid \log \boldsymbol{y}_r^0\rangle_1 \leqslant \log \overline{\delta}_r \Longleftrightarrow\\ &\quad\Longleftrightarrow \begin{cases}\left\langle \boldsymbol{\mu}^k \;\middle|\; \log \frac{1}{\boldsymbol{y}_r^0}\right\rangle_1 \leqslant \log \frac{1}{\underline{\delta}_r},\\ \langle \boldsymbol{\mu}^k \mid \log \boldsymbol{y}_r^0\rangle_1 \leqslant \log \overline{\delta}_r,\end{cases} \quad r = s^D + 1, \ldots, s^D + s^N.\end{aligned} \tag{3.4}$$

We do not impose any further unknown to the model, allowing the observations to be freely projected into the frontier as long as optima remain within the defined lower and upper bounds; and

(c) **Undesirable KPIs:**

$$\begin{aligned}&\langle \boldsymbol{\mu}^k \mid \boldsymbol{y}_r^-\rangle_0 \leqslant \min\{y_r^{k-}, \tau_r^-\} \Longleftrightarrow\\ &\quad\Longleftrightarrow \langle \boldsymbol{\mu}^k \mid \log \boldsymbol{y}_r^-\rangle_1 + Q_r^{k-} = \min\{\log y_r^{k-}, \log \tau_r^-\},\\ &\quad r = s^D + s^N + 1, \ldots, s,\end{aligned} \tag{3.5}$$

where Q_r^{k-} is a non-negative unknown quantity, and the equation holds from the convexity and the weak disposability related to undesirable criteria.

The non-negative unknowns Q_r^{k+} and Q_r^{k-} can be rewritten as the product of a known discretionary component, d_r^k, and a non-discretionary, radial, unconstrained, and unknown constant, η^k, for any $r = 1, \ldots, s$:

$$\begin{cases} Q_r^{k+} = d_r^k \eta^k,\ r = 1, \ldots, s^D,\\ Q_r^{k-} = d_r^k \eta^k,\ r = s^D + s^N + 1, \ldots, s,\end{cases} \tag{3.6}$$

being $\boldsymbol{d}^k = (d_1^k, \ldots, d_r^k, \ldots, d_s^k)$ an array defined by the user. Entries of $\boldsymbol{d}^k$ for $r = s^D + 1, \ldots, s^D + s^N$ (associated with the neutral KPIs) are meaningless in this framework, so the user does not have to define them. There are several ways of defining this array. Imposing it as equal to the observation of DMU k is a frequent alternative, but since we are in the logarithmic space of variables, $d_r^k = \log y_r^k$ seems a better choice. Thus, $Q_r^{k\pm} = \eta^k \log d_r^k$, $r = 1, \ldots, s$. Another alternative is the one proposed by Fusco (2015). If PC_r denotes the rth principal component of $\log \boldsymbol{Y}$, then $d_r^k = PC_r^k \cdot \frac{\mathrm{var}(PC_r)}{\mathrm{var}(PC_1)}$, where $\mathrm{var}(PC_r)$ stands for the variance of PC_r. This alternative mitigates the compensation effect that may still exist in our log-linear programming model.

Since all constraints above are linear, we may define a linear objective function to get a simple linear program to optimize parameters $\boldsymbol{\mu}^k$ and, eventually, the unknown constant η^k. Note that this parameter is the distance (in logarithmic terms) from the observation to the optimal values (or targets), directed by the array $\boldsymbol{d}^k$. Therefore, our model can simply find the maximum distance from the DMU k to the frontier; the objective function is, then, simply $(\eta^k)^* = \max \eta^k$, where the asterisk stands for the optimum. Provided that coefficients $\boldsymbol{\mu}^k$ must be non-negative and their sum should be one, we get the following linear programming problem:

$$(\eta^k)^* = \max_{\boldsymbol{\mu}^k, \eta^k}\ \eta^k \tag{3.7}$$

subject to:

$$\begin{cases}\left\langle \boldsymbol{\mu}^k \;\middle|\; \log \frac{1}{\boldsymbol{y}_r^+}\right\rangle_1 + d_r^k \eta^k \leqslant \min\left\{\log \frac{1}{y_r^{k+}},\ \log \frac{1}{\tau_r^+}\right\},\ r = 1, \ldots, s^D,\\ \left\langle \boldsymbol{\mu}^k \;\middle|\; \log \frac{1}{\boldsymbol{y}_r^0}\right\rangle_1 \leqslant \log \frac{1}{\underline{\delta}_r},\ r = s^D + 1, \ldots, s^D + s^N,\\ \langle \boldsymbol{\mu}^k \mid \log \boldsymbol{y}_r^0\rangle_1 \leqslant \log \overline{\delta}_r,\ r = s^D + 1, \ldots, s^D + s^N,\\ \langle \boldsymbol{\mu}^k \mid \log \boldsymbol{y}_r^-\rangle_1 + d_r^k \eta^k = \min\{\log y_r^{k-},\ \log \tau_r^-\},\ r = s^D + s^N + 1, \ldots, s,\\ \sum_j \mu_j^k = 1,\\ \boldsymbol{\mu}^k \geqslant 0.\end{cases} \tag{3.8}$$

Unfortunately, the constraint associated with the weak disposability of undesirable KPIs does not suffice to characterize the production process correctly (Kuosmanen, 2005; Kuosmanen & Podinovski, 2009). An array of abatement factors, $\boldsymbol{\zeta}^k = (\zeta_1^k, \ldots, \zeta_j^k, \ldots, \zeta_n^k)$ must be introduced into constraints , multiplying the coefficients $\boldsymbol{\mu}^k$ in the equation of undesirable KPIs:

$$\langle \boldsymbol{\mu}^k \odot \boldsymbol{\zeta}^k \mid \log \boldsymbol{y}_r^-\rangle_1 + d_r^k \eta^k = \min\{\log y_r^{k-}, \log \tau_r^-\},\ \text{for any } r = s^D + s^N + 1, \ldots, s, \tag{3.9}$$

where $\odot$ denotes the Hadamard's component-wise product between two arrays, *i.e.*,

$$\boldsymbol{\mu}^k \odot \boldsymbol{\zeta}^k = (\mu_1^k \zeta_1^k, \ldots, \mu_j^k \zeta_j^k, \ldots, \mu_n^k \zeta_n^k).$$

Because $\boldsymbol{\zeta}^k$ and $\boldsymbol{\mu}^k$ are both unknown (they are artifacts of the model, to be optimized), the new model becomes non-linear. Kuosmanen (2005) and Kuosmanen and Podinovski (2009) assume the existence of two additional non-negative unknown artifacts, $\boldsymbol{\sigma}^k = (\sigma_1^k, \ldots, \sigma_j^k, \ldots, \sigma_n^k)$ and

$\boldsymbol{\omega}^k = (\omega_1^k, \ldots, \omega_j^k, \ldots, \omega_n^k)$, such that:

$$\begin{cases} \boldsymbol{\mu}^k = \boldsymbol{\sigma}^k + \boldsymbol{\omega}^k, \\ \boldsymbol{\sigma}^k = \boldsymbol{\mu}^k \odot \boldsymbol{\zeta}^k, \\ \boldsymbol{\omega}^k = \boldsymbol{\mu}^k \odot (1 - \boldsymbol{\zeta}^k). \end{cases} \tag{3.10}$$

Therefore, the optimization of the parameter η^k results from the following linear model:

$$(\eta^k)^* = \max_{\boldsymbol{\sigma}^k, \boldsymbol{\omega}^k, \eta^k} \eta^k \tag{3.11}$$

subject to:

$$\begin{cases} \left\langle \boldsymbol{\sigma}^k \;\middle|\; \log \frac{1}{\boldsymbol{y}_r^+} \right\rangle_1 + \left\langle \boldsymbol{\omega}^k \;\middle|\; \log \frac{1}{\boldsymbol{y}_r^+} \right\rangle_1 + d_r^k \eta^k \\ \quad \leqslant \min\left\{ \log \frac{1}{y_r^{k+}}, \ \log \frac{1}{\tau_r^+} \right\}, \ r = 1, \ldots, s^D, \\ \left\langle \boldsymbol{\sigma}^k \;\middle|\; \log \frac{1}{\boldsymbol{y}_r^0} \right\rangle_1 + \left\langle \boldsymbol{\omega}^k \;\middle|\; \log \frac{1}{\boldsymbol{y}_r^0} \right\rangle_1 \\ \quad \leqslant \log \frac{1}{\underline{\delta}_r}, \ r = s^D + 1, \ldots, s^D + s^N, \\ \langle \boldsymbol{\sigma}^k \mid \log \boldsymbol{y}_r^0 \rangle_1 + \langle \boldsymbol{\omega}^k \mid \log \boldsymbol{y}_r^0 \rangle_1 \leqslant \log \overline{\delta}_r, \ r = s^D + 1, \ldots, s^D + s^N, \\ \langle \boldsymbol{\sigma}^k \mid \log \boldsymbol{y}_r^- \rangle_1 + d_r^k \eta^k = \min\{\log y_r^{k-}, \log \tau_r^-\}, \ r = s^D + s^N + 1, \ldots, s, \\ \sum_j \sigma_j^k + \omega_j^k = 1, \\ \boldsymbol{\sigma}^k \geqslant 0, \\ \boldsymbol{\omega}^k \geqslant 0. \end{cases} \tag{3.12}$$

Now, thanks to the optima of $\boldsymbol{\sigma}^k$ and $\boldsymbol{\omega}^k$ from Eq. (3.12), we can use the following metric to construct a *performance* score associated with DMU k:

$$\theta^k = \prod_{r=1}^{s^D} \left(\frac{(y_r^{k+})^*}{y_r^{k+}} \right)^{-1/s^D} \cdot \prod_{r=s^D+1}^{s^D+s^N} \left(\frac{(y_r^{k0})^*}{y_r^{k0}} \right)^{\lambda_r^k / s^N} \cdot \prod_{r=s^D+s^N+1}^{s} \left(\frac{(y_r^{k-})^*}{y_r^{k-}} \right)^{1/s^U}, \tag{3.13}$$

where:

$$\lambda_r^k = \begin{cases} 1, \ \text{if } y_r^{k0} \geqslant (y_r^{k0})^*, \\ -1, \ \text{otherwise}, \end{cases} \quad r = s^D + 1, \ldots, s^D + s^N, \tag{3.14}$$

and:

$$\begin{cases} (y_r^{k+})^* = \exp\langle \boldsymbol{\sigma}^k + \boldsymbol{\omega}^k \mid \log \boldsymbol{y}_r^+ \rangle_1 = \exp \sum_j (\sigma_j^k + \omega_j^k) \log y_r^{j+}, \ r = 1, \ldots, s^D, \\ (y_r^{k0})^* = \exp\langle \boldsymbol{\sigma}^k + \boldsymbol{\omega}^k \mid \log \boldsymbol{y}_r^0 \rangle_1 = \exp \sum_j (\sigma_j^k + \omega_j^k) \log y_r^{j0}, \ r = s^D + 1, \ldots, s^D + s^N, \\ (y_r^{k-})^* = \exp\langle \boldsymbol{\sigma}^k \mid \log \boldsymbol{y}_r^- \rangle_1 = \exp \sum_j \sigma_j^k \log y_r^{j-}, \ r = s^D + s^N + 1, \ldots, s, \end{cases} \tag{3.15}$$

which holds from Eq. (3.2). One should interpret this score as the *Composite Indicator* associated with DMU k and all s KPIs.

Finally, we have the optimal value θ^k as the performance score of DMU k. If $\theta^k = 1$, then k is a top performer (benchmark) and is located in the frontier; otherwise, $\theta^k < 1$ and there are sources hampering the good performance. We can use this performance score to rank DMUs in descending order. Benchmarks have higher scores and higher ranks. At the limit, all DMUs can be considered benchmarks as they can all be positioned in the frontier, having $\theta^k = 1$; the converse is not true as at least one DMUs must be a benchmark. Please, refer to Example 3.1, below, which shows: (*i*) the imposition of ranges for neutral KPIs; (*ii*) the imposition of regulatory thresholds for both desirable and undesirable KPIs; (*iii*) the assessment of targets, or optimal values, per KPI and DMU; and (*iv*) the aggregated performance score per DMU concerning the targets and the observations as did by Eq. (3.13).

Example 3.1. Let us consider a set of ten DMUs, as well as two desirable, one undesirable, and one neutral KPIs. Table 3.1(*i*) contains the data and the limits for the neutral KPI per DMU (note that those limits can be DMU-specific). Consider that $\boldsymbol{d}^j = \log \boldsymbol{y}^j = (\log y_1^j, \ldots, \log y_4^j)$ for any $j = 1, \ldots, 10$. We applied Model (3.12) to optimize the targets associated with the four KPIs. Fixing $\tau_1^+ = \tau_2^+ = 75$ and $\tau_4^- = 100$ means that the DMUs should have $(y_1^+)^*$ and $(y_2^+)^* \geqslant 75$ and, simultaneously, $(y_4^-)^* \leqslant 100$ to be considered benchmark. Additionally, neutral KPIs should obey to the bottom and top levels defined in the last two columns of Table 3.1(*i*). Table 3.2 exhibits the optimized coefficients $\boldsymbol{\sigma}^k$ and $\boldsymbol{\omega}^k$ per DMU. It is interesting to note that, according to this model, DMUs can contribute to construct a benchmark but might not be benchmarks themselves. For instance, DMU 7 contributes to achieve the targets of DMU 1 with $\mu_7^1 = \sigma_7^1 + \omega_7^1 = 0.0992 + 0.0589 = 0.1581$ (for desirable and neutral KPIs) and $\sigma_7^1 = 0.0992$ (for undesirable KPIs). Therefore, according to Eq. (3.15), the targets of DMU 1 are as follows:

$$\begin{cases} (y_1^{1+})^* = \exp\{(0.5666 + 0.1897) \log 100 + (0.0992 + 0.0589) \log 25 + (0.0267 + 0.0588) \log 50\} \\ \quad = 75.7 \ (\geqslant 75), \\ (y_2^{1+})^* = \exp\{(0.5666 + 0.1897) \log 100 + (0.0992 + 0.0589) \log 125 + (0.0267 + 0.0588) \log 75\} \\ \quad = 101.1 \ (\geqslant 75), \\ (y_3^{1,0})^* = \exp\{(0.5666 + 0.1897) \log 75 + (0.0992 + 0.0589) \log 25 + (0.0267 + 0.0588) \log 5\} \\ \quad = 50 \ (\in [0, 50]), \\ (y_4^{1-})^* = \exp\{0.5666 \log 100 + 0.0992 \log 125 + 0.0267 \log 100\} \\ \quad = 24.8 \ (\leqslant 100). \end{cases}$$

Table 3.1(*ii*) provides the targets for DMUs 1 up to 10. Comparing the averages of observations and targets, we verify the need to increase desirable KPIs by 41%–45%, a smaller change of 9% in neutral KPIs, and a decrease of the undesirable KPI by 14% (on average). Table 3.1(*iii*) contains the values of η^k and θ^k (performance score). To compute the performance score, θ^k for $k = 1$, we use those targets and compare them to observations:

$$\begin{cases} (y_1^{1+})^* / y_1^{1+} = 75.7/50 = 1.514, \\ (y_2^{1+})^* / y_2^{1+} = 101.1/100 = 1.011, \\ (y_3^{1,0})^* / y_3^{1,0} = 50/50 = 1, \\ (y_4^{1-})^* / y_4^{1-} = 24.8/25 = 0.992, \\ \text{which imply} \\ \theta^1 = 0.992 \cdot 1/\sqrt{1.514 \cdot 1.011} = 0.802. \end{cases}$$

These results show that DMU 1 should raise KPI 1 by 51% and KPI 2 by just 1%, while holding KPI 3 and reducing KPI 4 by only 0.8%. It is interesting to note that $(\eta^k)^* = 0$ does not imply $\theta^k = 1$; thus, we should not use the former for DMUs' ranking. □

Example 3.2. Let us consider the same data as in Table 3.1(*i*), except for τ_1^+, τ_2^+, τ_4^-, $\underline{\delta}_3$, and $\overline{\delta}_3$. In this case, we assume that regulators and other decision-making players are less restrictive, imposing $\tau_1^+ = \tau_2^+ = 0$ and $\tau_4^- = +\infty$. Additionally, there are no bounds for the neutral KPI, which means that it is free to change (increase or decrease or even remain) to enhance performance: $\underline{\delta}_3 = 0$ and $\overline{\delta}_3 = +\infty$. Indeed, back to Eq. (3.4), we have now $\langle \boldsymbol{\mu}^k \mid \boldsymbol{y}_r^0 \rangle_0 \geqslant 0 \ \wedge \ \langle \boldsymbol{\mu}^k \mid \boldsymbol{y}_r^0 \rangle_0 \leqslant +\infty$, which is a tautology provided that $y_r^0 > 0$ and $\boldsymbol{\mu}^k \geqslant 0$. Nonetheless, this neutral KPI still counts to the performance assessment because its associated target-to-observation ratio is not necessarily unitary, thus having to be accounted for in Eq. (3.13). Table 3.3 contains the main results associated with this iteration. Because the model is now less

Table 3.1
Example's data and results obtained using Eq. (3.12) for optimization.

(*i*) Data

DMU	y_1^+	y_2^+	y_3^0	y_4^-	$\underline{\delta}_3$	$\overline{\delta}_3$
1	50	100	50	25	0	50
2	75	25	125	50	0	50
3	100	75	100	75	0	50
4	100	100	75	100	25	75
5	50	50	50	25	25	75
6	75	50	50	75	25	75
7	25	125	25	125	50	100
8	50	75	5	100	50	100
9	25	25	100	25	75	125
10	100	50	25	50	75	125
Mean	65.0	67.5	60.5	65		
$\tau_r^{\pm}$	75	75	–	100	–	–

(*ii*) Targets

DMU	$(y_1^+)^*$	$(y_2^+)^*$	$(y_3^0)^*$	$(y_4^-)^*$
1	75.7	101.1	50.0	24.8
2	93.7	88.5	50.0	40.9
3	100.0	76.0	48.9	75.0
4	100.0	100.0	75.0	100.0
5	100.0	100.0	75.0	19.7
6	100.0	97.7	73.2	56.3
7	75.0	105.3	58.1	100.0
8	97.7	100.4	73.6	73.3
9	100.0	100.0	75.0	18.7
10	100.0	83.7	83.1	50.0
Mean	94.2	95.3	66.2	55.9
Δ (%)	44.9	41.1	9.4	−14.0

(*iii*) Performance

DMU	$(\eta^k)^*$	θ^k
1	−0.002	0.802
2	−0.052	0.155
3	0	0.485
4	0	1.000
5	−0.074	0.263
6	−0.067	0.317
7	0.033	0.228
8	−0.068	0.031
9	−0.089	0.141
10	0	0.232
Mean		0.366
Std. Dev.		0.294

Table 3.2
Optimal coefficients $\boldsymbol{\sigma}^k$ and $\boldsymbol{\omega}^k$[a].

DMU	$\sigma_j^k,\ j=1,\ldots,10$									
	1	2	3	4	5	6	7	8	9	10
1				0.5666			0.0992	0.0267		
2				0.6358				0.0828		0.1026
3			0.0087	0.5979						0.3903
4				1.0000						
5				0.6476						
6			0.0177	0.8375						0.0247
7				0.7675			0.2325			
8				0.9272			0.0051			
9				0.6365						
10			0.4402	0.4093						0.0324
DMU	$\omega_j^k,\ j=1,\ldots,10$									
	1	2	3	4	5	6	7	8	9	10
1				0.1897			0.0589	0.0588		
2				0.1332				0.0115		0.0342
3			0.0009	0.0012						0.0011
4										
5				0.3524						
6			0.0005	0.1180						0.0016
7										
8				0.0558			0.0119			
9				0.3635						
10			0.0763	0.0325						0.0093

[a]Empty cells indicate that optimal coefficients are zero.

restrictive, the average performance score is larger than the observed in Table 3.1(*iii*) for Example 3.1. There are two DMUs with maximum (unitary) performance score (DMUs 4 and 7), instead of just one as happened before. However, performance scores can decrease as with DMUs 1, 5, and 9. As in Example 3.1, DMUs 3, 4, 7, and 10 act as benchmarks for the construction of targets through non-null coefficients $\boldsymbol{\sigma}^k$ and $\boldsymbol{\omega}^k$, although only 4 and 7 exhibit $\theta^k = 1$. □

4. The composite indicator when some data are imperfectly known

In some cases data are imperfectly known because of diverse reasons, including data missing and measurement errors (French, 1995; Hayek, 1945; Roy et al., 2014). The listwise deletion, *i.e.*, removing entities that do not provide some data, may introduce severe problems into the BoD model. Indeed, the latter suffers from the so-called *curse of dimensionality* problem. Putting it differently, the model is sensitive to the sample's size, being the removal of a substantial quantity of DMUs problematic as the results' resolution may be lost. Several techniques have been proposed to overcome such an issue. Perhaps the most employed one is the imputation of values to replace imperfectly known or missing data (Olinsky et al., 2003; Raaijmakers, 1999; Reilly, 1993). Imputation or substitution can be useful if reasonable estimates can be achieved to be substitutes. However, estimates may not be robust enough to guarantee satisfactory outcomes in performance assessment. It is because of the difficulty of ensuring that the model used for estimation (*e.g.*, a regression) categorically fits the reality. Even if that model obeys certain assumptions, it may be impossible to confirm that the estimate is close to the real observation. However, imputation has its own merits, including not reducing the sample's size.

In any case, imposing boundaries for missing or imperfectly known data emerges as a reasonable alternative to fix a single value for imputation (Ferreira, Figueira, et al., 2021). It reduces the difficulty of defining suitable estimates, being enough to state lower and upper limits. Of course, ill-defined or too broad intervals can be noxious for the quality of outcomes. Suppose all data for a given DMU be perfectly known. In that case, the defined interval reduces itself to a single point in the Euclidean space (the upper and the lower bounds are equal to

Table 3.3
Results associated with the Example 3.2.

DMU	$(y_1^+)^*$	$(y_2^+)^*$	$(y_3^0)^*$	$(y_4^-)^*$
1	54.35	110.31	46.26	23.34
2	100.00	78.86	56.23	38.53
3	100.00	90.61	68.65	75.00
4	100.00	100.00	75.00	100.00
5	100.00	100.00	75.00	14.13
6	100.00	89.28	67.65	56.25
7	25.00	125.00	25.00	125.00
8	68.56	106.26	55.61	68.96
9	100.00	100.00	75.00	6.25
10	100.00	72.69	51.51	50.00
Mean	84.79	97.30	59.59	55.75
Std.Dev.	25.27	14.47	15.25	35.86

DMU	$\frac{(y_1^+)^*}{y_1^+}$	$\frac{(y_2^+)^*}{y_2^+}$	$\frac{(y_3^0)^*}{y_3^0}$	$\frac{(y_4^-)^*}{y_4^-}$
1	1.09	1.10	0.93	0.93
2	1.33	3.15	0.45	0.77
3	1.00	1.21	0.69	1.00
4	1.00	1.00	1.00	1.00
5	2.00	2.00	1.50	0.57
6	1.33	1.79	1.35	0.75
7	1.00	1.00	1.00	1.00
8	1.37	1.42	11.12	0.69
9	4.00	4.00	0.75	0.25
10	1.00	1.45	2.06	1.00
Mean	1.51	1.81	2.08	0.80
Std.Dev.	0.88	0.95	3.04	0.23

DMU	$(\eta^k)^*$	θ^k
1	−0.021	0.789
2	0	0.169
3	0	0.625
4	0	1.000
5	−0.177	0.188
6	−0.067	0.359
7	0	1.000
8	0	0.044
9	−0.431	0.047
10	0	0.403
Mean		0.462
Std.Dev.		0.351

the observation). Otherwise, a broader interval is desirable to define the possible domain of the imperfect observation.

Following (Despotis & Smirlis, 2002) and Smirlis et al. (2006), we may define the following generic intervals, being l_r^k and u_r^k, respectively,[3] the lower and upper bounds associated with the rth variable and DMU k:

(a) $y_r^{k+} \in [l_r^{k+}, u_r^{k+}]$, $r = 1, \dots, s^D$, for desirable KPIs;
(b) $y_r^{k0} \in [l_r^{k0}, u_r^{k0}]$, $r = s^D, \dots, s^D + s^N$, for neutral KPIs; and
(c) $y_r^{k-} \in [l_r^{k-}, u_r^{k-}]$, $r = s^D + s^N + 1, \dots, s$, for undesirable KPIs.

The construction of these intervals is out of the scope of this paper, but we refer the reader to the work of Cherchye et al. (2011) that shows how to build them based on multiple imputation. In some other cases, experts can define those intervals (Ferreira, Marques, & Pedro, 2018).

[3] In the case of perfect knowledge of data for DMU k and KPI r: $l_r^k = u_r^k$.

Now, let us consider two distinct scenarios (Ferreira & Marques, 2021):

Worst scenario. $\mathcal{S}^W = \{(y_r^+, y_r^0, y_r^-) | y_r^+ = l_r^+,\ y_r^0 \in [l_r^0, u_r^0],\ y_r^- = u_r^-\}$, *i.e.*, data are in the most unfavorable situation (undesirable KPIs are maximal whereas the desirable ones are minimal according to the intervals defined before).

Best scenario. $\mathcal{S}^B = \{(y_r^+, y_r^0, y_r^-) | y_r^+ = u_r^+,\ y_r^0 \in [l_r^0, u_r^0],\ y_r^- = l_r^-\}$, as opposed to the worst scenario.

It is easy to conclude that the frontier estimated using $\mathcal{S}^B$, denoted ∂F^B, outperforms the frontier associated with $\mathcal{S}^W$, *i.e.*, ∂F^W. Because of these scenarios, we can obtain an interval for the performance θ^k associated with DMU k. Indeed, by projecting the array of entries $(y_r^{k+}, y_r^{k0}, y_r^{k-}) \in \mathcal{S}^B$ into the frontier ∂F^W through Eqs. (3.11)–(3.12), we achieve the smallest performance estimate $\underline{\theta}^k$. In the same vein, the maximum distance of k to the frontier, given uncertainty or missing data, is $\overline{\theta}^k$ and results from projecting $(y_r^{k+}, y_r^{k0}, y_r^{k-}) \in \mathcal{S}^W$ into ∂F^B. Hence, it should be clear that $\theta^k \in [\underline{\theta}^k, \overline{\theta}^k]$ (Despotis & Smirlis, 2002; Smirlis et al., 2006).

Although useful to establish boundaries for the distance (and performance) of DMU k to the frontier, this approach does not give us the *whole picture* about the performance of that DMU. Indeed, provided that some data are imperfectly known or even totally unknown, the process of estimating performance levels may become partially stochastic. Thus, there is an unknown probability associated with the scenario "*DMU k is at the distance η^k to the frontier and, for that reason, has a performance score θ^k*". In practice, we do not know these odds; so, we need to simulate the production process by implementing a Monte-Carlo simulation routine. Algorithm 1 presents it.

Algorithm 1 starts by fixing the boundaries per KPI and DMU, as well as DMU k of which one desires to estimate the performance. Another input of this Monte-Carlo algorithm is the number of iterations, T, which is typically in the thousands, say $T = 5000$ iterations. In each iteration, the algorithm draws randomly and uniformly a single quantity from each interval fixed in the pre-processing. This process is carried out for each DMU in the set. By assembling those random (but bounded) values, one constructs the new matrix $\boldsymbol{Y}^{(t)}$, that might differ from the original matrix of data, $\boldsymbol{Y}$. By using Eq. (3.11) and the new matrix $\boldsymbol{Y}^{(t)}$ to construct the new frontier, $\partial F^{(t)}$, associated with the tth iteration, one estimates a performance score for DMU k and iteration t, *i.e.*, $(\theta^k)^{(t)}$. By the end of the Monte-Carlo cycle, one has T performance estimates, sorted in the array θ^k, all associated with DMU k; see Example 4.1. One can perform multiple operations using this array: obtaining basic statistics (average, standard deviation, …) is perhaps the most fundamental one. An important indicator is $F(\theta^j \geqslant \xi)$, *i.e.*, the probability of θ^j being at least equal to a certain threshold $\xi \in [0, 1]$. If ξ approaches 1, say $\xi > 0.95$, the DMU gets close to the frontier. Therefore, we can build upon (Cherchye et al., 2011) to classify DMUs as follows:

1. *Benchmarks* – DMUs classified as benchmarks exhibit $F(\theta^j \geqslant \xi) = 100$ for $\xi \to 1$;
2. *Potential benchmarks* – these DMUs exhibit $0 < F(\theta^j \geqslant \xi) < 100$ for $\xi \to 1$, *i.e.* there is a chance that they can turn into benchmarks;
3. *DMUs open to improvement* – DMUs demanding improvement are those falling apart the frontier regardless of data, *i.e.*, $F(\theta^j \geqslant \xi) = 0$ for $\xi \to 1$.

Example 4.1. Let us consider once more the case of Example 3.2. We verified before that DMU $k = 4$ was considered a benchmark, $\theta^k = 1$. Suppose now that its data were considered imperfectly known because of errors on measuring instruments. Thus, we must impose boundaries to the observations of this DMU. For this example, let us consider the following boundaries: $[l_1^{k+}, u_1^{k+}] = [50, 150]$, $[l_2^{k+}, u_2^{k+}] =$

Algorithm 1 Monte-Carlo simulation routine for imperfect knowledge of data modeling

Input $[l_r^{j+}, u_r^{j+}]$, $r = 1, \dots, s^D$, $j = 1, \dots, n$
Input $[l_r^{j0}, u_r^{j0}]$, $r = s^D + 1, \dots, s^D + s^N$, $j = 1, \dots, n$
Input $[l_r^{j-}, u_r^{j-}]$, $r = s^D + s^N + 1, \dots, s$, $j = 1, \dots, n$
Input k
Input T
1: **procedure** MONTE-CARLO
2: **for** $t = 1$ to $t = T$ **do**
3: Define an empty array, $\theta^{(t)} \leftarrow \{\}$;
4: Define an empty $n \times s$ matrix, $\boldsymbol{Y}^{(t)} \leftarrow \{\}_{n\times s}$;
5: **for** $j = 1$ to $j = n$ **do**
6: Randomly select any value $(y_r^{j+})^{(t)} \in [l_r^{j+}, u_r^{j+}]$, $r = 1, \dots, s^D$;
7: Randomly select any value $(y_r^{j0})^{(t)} \in [l_r^{j0}, u_r^{j0}]$, $r = s^D + 1, \dots, s^D + s^N$;
8: Randomly select any value $(y_r^{j-})^{(t)} \in [l_r^{j-}, u_r^{j-}]$, $r = s^D + s^N + 1, \dots, s$;
9: Construct the array $(y^j)^{(t)} = ((y_r^{j+})^{(t)}, (y_r^{j0})^{(t)}, (y_r^{j-})^{(t)})^\top$;
10: Construct the matrix $\boldsymbol{Y}^{(t)} \leftarrow \{\boldsymbol{Y}^{(t)}, (y^j)^{(t)}\}$;
11: **end for**
12: Use Eq. (3.11) and matrix $\boldsymbol{Y}^{(t)}$ to project $(y^k)^{(t)}$ into $\partial F^{(t)}$ and obtain $(\theta^k)^{(t)}$;
13: Update the array $\theta^k \leftarrow \{\theta^k, \mathbb{E}[(\theta^k)^{(t)}]\}$.
14: **end for**
15: Sort the array θ^k in ascending order.
16: **end procedure**
Output θ^k

$[75, 125]$, $[l_3^{k0}, u_3^{k0}] = [0, 100]$, $[l_4^{k-}, u_4^{k-}] = [25, 175]$. Data associated with the other nine DMUs were considered perfectly known, thus the intervals reduce themselves to a single point (the real observation). Additionally, let $T = 10$ for simplicity. As it is well known, when one observation changes, the performance of the remaining DMUs is likely to change as well. Therefore, we have to access the performance of all DMUs in the set. It means to apply Algorithm 1 ten times, once per DMU. However, we can take advantage of what was done in rows 5–11 and simplify the routine, making it faster. It is specially relevant in large problems, with lots of DMUs. Table 4.1 presents the randomly generated values for DMU $k = 4$ and $T = 10$ iterations. Using these data, we can construct the matrix $\boldsymbol{Y}^{(t)}$ required for each iteration (rows 10 and 12 of Algorithm 1). For instance, concerning the iteration $t = 9$, the data matrix is

$$\boldsymbol{Y}^{(9)} = \begin{pmatrix} 50 & 100 & 50 & 25 \\ 75 & 25 & 125 & 50 \\ 100 & 75 & 100 & 75 \\ 98.24 & 102.52 & 22.29 & 60.34 \\ 50 & 50 & 50 & 25 \\ 75 & 50 & 50 & 75 \\ 25 & 125 & 25 & 125 \\ 50 & 75 & 5 & 100 \\ 25 & 25 & 100 & 25 \\ 100 & 50 & 25 & 50 \end{pmatrix}.$$

We can see the results of the ten iterations in Table 4.2, as well as the initial performance estimates (as in Example 3.2), and the average difference to them, Δ (%). This table clearly shows that perturbations in a single DMU can lead to considerable changes in the performance of others. In this example and after just ten iterations, such a perturbation reached over 40% compared to the initial performance estimate (DMU 3). DMU 4, which was initially considered a benchmark, was not always like that (in two iterations, or 20% of the cases, the estimated performance was low). In opposition, DMU 7 was always a benchmark, regardless of the changes that happened in DMU 4.

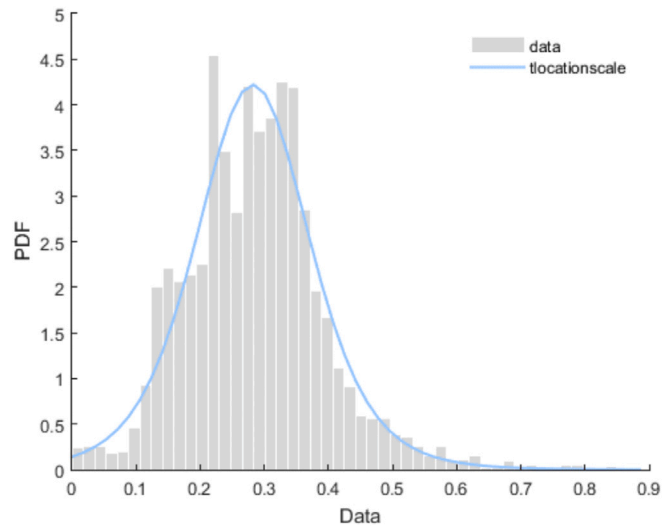

Fig. 4.1. Empirical and theoretical distributions of DMU 10's performance (Example 4.1), after 5000 iterations of Algorithm 1.

Redoing the same exercise with $T = 5000$ iterations, we obtained the results displayed in Table 4.3, where CV stands for the coefficient of variation and $\text{F}(\theta^j \geqslant 0.95)$ for the probability of DMU j exhibiting $\theta^j \geqslant 0.95$, i.e., being close to or positioned in the frontier. Besides, we classified entities as *benchmarks* if $\text{F}(\theta^j \geqslant 0.95) = 100$ as $\xi = 0.95$ is close to 1, *potential benchmarks* if $\text{F}(\theta^j \geqslant 0.95)$ was between 0 and 100 (given $\xi = 0.95$), and *open to improvement* if $\text{F}(\theta^j \geqslant 0.95)=0$. Indeed, for the latter, there is no chance these observations could be observed in the frontier; so, they are inefficient, falling apart the frontier. Meanwhile, there is a certain likelihood that entitled classified as *potential benchmarks* be classified as *benchmarks*, as all depends on the knowledge about the data quality.

These results confirm the previous ones: DMU 7 was a benchmark, being this fact independent from the data variations in DMU 4, which, in turn, was close to the frontier (or located there) in 95% of the iterations, being a potential benchmark. Others can reach unitary performance scores, but their average performance score is low (DMUs 1 and 3). The remaining DMUs cannot be benchmarks, being, thus, open to improvement.

Additionally, and as expected, changes in the data of a single DMU result, in general, in large deviations in the performance of the other DMUs. It can be confirmed by the considerable coefficients of variation in most of the DMUs. Let us take a look into DMU 10. Fig. 4.1 portrays the histogram of its performance estimates, as well as the theoretical probability density function that fits it the best (based on the Akaike Information Criterion). In the present case, such a function is a t location-scale with parameters 0.2821 (location), 0.0909 (scale), and 6.3549 (shape). When the shape parameter is sufficiently large, the t location-scale distribution function approaches the Gaussian distribution function (Elster & Klauenberg, 2020). As a matter of fact, the Gaussian density that fits this empirical distribution has parameters 0.2867 (mean or location) and 0.1099 (standard deviation or scale), which are close to the location and scale parameters of the t location-scale function. Therefore, we can make statistical inference, including statistical tests. For instance, the probability of the performance of DMU 10 being larger than a value ζ is $1-\phi\left(\frac{\zeta-0.2867}{0.1099}\right)$, where ϕ stands for the

Table 4.1
Randomly generated observations for DMU $k = 4$ and ten iterations.

	$t = 1$	$t = 2$	$t = 3$	$t = 4$	$t = 5$	$t = 6$	$t = 7$	$t = 8$	$t = 9$	$t = 10$
$(y_1^{k+})^{(t)}$	52.37	55.28	57.63	71.87	56.39	120.64	91.47	132.73	98.24	81.66
$(y_2^{k+})^{(t)}$	77.56	119.69	109.83	95.18	107.52	117.87	113.01	100.23	102.52	76.82
$(y_3^{k0})^{(t)}$	35.04	30.89	30.28	53.55	63.19	66.44	40.20	60.57	22.29	12.29
$(y_4^{k-})^{(t)}$	44.05	148.57	31.60	126.70	166.77	140.54	161.19	72.47	60.34	60.80

Table 4.2
Results of ten iterations of Algorithm 1.

	Initial	$t = 1$	$t = 2$	$t = 3$	$t = 4$	$t = 5$	$t = 6$	$t = 7$	$t = 8$	$t = 9$	$t = 10$	Average	Δ (%)
$j = 1$	0.789	1	0.511	0.506	0.795	0.734	0.624	0.523	0.629	0.392	1	0.671	−14.90%
$j = 2$	0.169	0.164	0.217	0.164	0.217	0.178	0.125	0.276	0.108	0.163	0.217	0.183	8.31%
$j = 3$	0.625	1	1	1	1	1	0.405	1	0.349	1	1	0.875	40.16%
$j = 4$	1	0.422	1	1	1	1	1	1	1	1	0.133	0.855	−14.45%
$j = 5$	0.188	0.251	0.261	0.279	0.263	0.225	0.160	0.262	0.250	0.157	0.251	0.236	25.10%
$j = 6$	0.359	0.285	0.348	0.291	0.391	0.305	0.240	0.277	0.248	0.345	0.394	0.313	−12.99%
$j = 7$	1	1	1	1	1	1	1	1	1	1	1	1	0.00%
$j = 8$	0.044	0.057	0.059	0.059	0.059	0.041	0.039	0.057	0.054	0.105	0.057	0.059	31.56%
$j = 9$	0.047	0.077	0.063	0.061	0.064	0.077	0.029	0.041	0.031	0.017	0.077	0.054	14.40%
$j = 10$	0.403	0.310	0.317	0.324	0.535	0.527	0.190	0.222	0.199	0.335	0.384	0.334	−16.95%

Table 4.3
Main statistics of performance after 5000 iterations of Algorithm 1.

	Initial	Average	Std.Dev.	CV	min	max	$F(\theta^j \geqslant 0.95)$	Classification
$j = 1$	0.789	0.623	0.249	40%	0.004	1.000	15%	Potential benchmark
$j = 2$	0.169	0.150	0.072	48%	0.000	0.305	0%	Open to improvement
$j = 3$	0.625	0.754	0.344	46%	0.000	1.000	63%	Potential benchmark
$j = 4$	1	0.978	0.115	12%	0.005	1.000	95%	Potential benchmark
$j = 5$	0.188	0.205	0.076	37%	0.000	0.404	0%	Open to improvement
$j = 6$	0.359	0.273	0.094	34%	0.000	0.496	0%	Open to improvement
$j = 7$	1	1	0	0%	1.000	1.000	100%	Benchmark
$j = 8$	0.044	0.080	0.068	85%	0.008	0.600	0%	Open to improvement
$j = 9$	0.047	0.043	0.024	56%	0.000	0.092	0%	Open to improvement
$j = 10$	0.403	0.287	0.110	38%	0.000	0.885	0%	Open to improvement

cumulative distribution function of the standard normal distribution. Should $\zeta = 0.95$, that probability becomes $1 - \phi(6.035) \approx 1 - 1 = 0$, as expected. □

5. Case study: The portuguese solid urban waste collection utilities

5.1. The waste sector in Portugal

Urban or municipal solid waste (MSW) includes waste collected by either (or on behalf of) municipal authorities or the private sector (business or private non-for-profit). Waste can be recyclable or not, and can be classified depending on its destination. The designation of MSW is still commonly used by different authors to differentiate urban waste from liquid or semi-liquid waste such as domestic wastewater. In Portugal, although the wastewater service is interconnected with the urban waste usually provided by the same municipal utility, the MSW terminology does not include wastewater (Marques & Simões, 2009).

In Europe, the legislation specifies municipal waste as (a) "*mixed waste and separately collected waste from households, including paper and cardboard, glass, metals, plastics, bio-waste, wood, textiles, packaging, waste electrical and electronic equipment, waste batteries and accumulators, and bulky waste, including mattresses and furniture*"; and (b) "*mixed waste and separately collected waste from other sources, where such waste is similar in nature and composition to waste from households*" (Directive 2008/98/EC of the European Parliament).

Over the past decades, Portuguese policies around MSW collection management have undergone rapid changes. According to Ferreira et al. (2020) "*these changes can be blamed for the serious environmental dysfunctions, the scarcity of appropriate waste disposal sites, the rising costs of collection and treatment systems, and even society's increased awareness of environmental issues* [...] *Strategies to encourage lower waste generation meant that the waste producer should properly collect, store, transport, dispose of and use waste without endangering human health or causing harm to the environment.*"

In Portugal, collection of municipal waste is guaranteed by different municipal utilities that comprise organized networks of human resources, logistics, equipment, and infrastructure necessary to carry out the operations inherent to the management of MSW. Waste management systems are segmented into two categories based on their activities: 'low' or 'retail' (collection from disposal sites to transfer stations/treatment plants) and 'high' or 'wholesaler' services (from transfer stations to landfill or other treatment/disposal process). Most of the 255 retail services (93%) are directly managed by the municipality they serve, and the remaining 7% have their management delegated to municipal or inter-municipal companies. In opposition, the management of the 23 wholesale services is mostly under municipal concessions (52%) or delegated to municipal or inter-municipal companies (35%). Retail services are responsible for collecting refuse waste (no sort of separation and prior selection of waste), while wholesale

Table 5.1
List of considered dimensions.

Criteria	Sub-criteria	Indicators	Description
[C1] Adequacy of infrastructures	[SC1.1] Accessibility	[RU01b] Physical accessibility of the service (%)	Number of households served with refuse waste collection service per 100 houses
		[RU02ab] Accessibility of the selective collection service (%)	Number of households served with selective waste collection service per 100 houses
		[RU03b] Economic accessibility of the service (%)	Average burden associated with the municipal waste management service divided by average household disposable income (×100)
	[SC1.2] Quality	[RU04b] Container cleaning (%)	Number of washes of containers for refuse waste deposit per 100 containers (surface or underground)
[C2] Management sustainability	[SC2.1] Cost recovery	[RU06ab] Coverage of expenses (%)	Utilities' income (tariffs and investment subsidies) per 100€of operational expenditures
	[SC2.2] Infrastructure	[RU07b] Recycling of selective waste collection (%)	Quantity of collected waste for recycling divided by the annual target of selective waste collection (×100)
		[RU11b] Vehicle renewal (km/vehicle)	Cumulative miles associated with waste collection per vehicle (selective or refuse waste collection)
		[RU12b] Monetization of vehicles (kg/m^3)	Quantity of urban waste indiscriminately collected divided by the installed capacity of vehicles for waste collection
	[SC2.3] Human resources productivity	[RU13b] Adequacy of human resources (no/1000ton)	Staff allocated to waste management (including outsourcing) per 1,000 tons of municipal collected waste
[C3] Environment sustainability	[SC3.1] Natural resources exploitation	[RU14b] Use of energy resources (tep/ton)	Consumed fuel associated with the refuse waste collection per ton of urban refuse waste
	[SC3.2] Pollution prevention	[RU16b] Emission of greenhouse gases from the refuse waste collection (kg CO_2/ton)	CO_2 emissions by vehicles of refuse waste collection per tone of urban refuse waste

services are responsible for collecting selective waste (although retail services could also manage the process).

MSW management are public and must be ruled by the universality, continuity, quality, efficiency, and equity principles. In Portugal, the sector of water and waste is regulated since 2009 by the Water and Waste Services Regulation Authority (ERSAR), by Decree-Law No. 277/2009 of October 2 (previously the urban waste was regulated by IRAR — the institute for regulation of water and waste services). The entity aims at contributing to a better sector organization, monitoring and periodically reporting on best practices to increase the services' efficiency, and sanctioning poor performance cases. ERSAR acts to prevent pernicious practices resulting from the legal monopolies (instead of natural monopolies), like inflated or inefficient prices, as typically each geographical area has just one service provider. Therefore, the end-user cannot choose between two operators, leading to the absence of competition.

5.2. Sample and treatment of data

In this case study, we considered only the retail services. Wholesale services were disregarded as they constitute a small sample (compared to the retail services) and, besides, each covers more than one municipality, meaning that other effects like scale and scope economies need to be analyzed, which is not aim of this paper; see the works of Caldas et al. (2019), Carvalho and Marques (2014), Simões et al. (2013). We initially retrieved the 2019-related data for the 253 retail services as made available officially and publicly by ERSAR. However, 9.1% of the sample was composed of observations whose data were absent for most of KPIs (*vide infra*) or the regulator itself considered that the data quality was poor. Even if we introduced Algorithm 1 to solve this problem of imperfect knowledge of data, we could not expect good outputs given the high variability we would introduce into the analysis. Therefore, we opted by removing those observations from the sample, ending up with 230 retail services, whose performance was assessed with the models presented before. These were clustered in terms of the intervention area typology as follows: urban areas (8.7%), rural areas (65.2%), and neither urban nor rural, also known as moderately urbanized areas (26.1%).

The considered sample of 232 observations was not featured by perfect knowledge of data. Indeed, there were some blank entries as well as some values whose quality was doubtful. As long as the number of variables with imperfect knowledge of data was, at the most, two we kept the observation and applied Algorithm 1. In these cases, we investigated whether the service was located in a rural, urban, or neither of both. Then, we defined the interval width using the standard deviation of the subsample to which the observation belonged. However, some lower and upper levels needed to be corrected as we imposed that they should not be below or above the minimum or the maximum (respectively) of the subsample. This procedure helped us to avoid negative or nonsense simulated values (for instance, percentages above 100%). For the cases of blank entries, we estimated the centroid of the interval with simple extrapolation with a linear regression and a perfectly known variable as independent. There were 308 (19%) entries that required these procedures.

All computations were performed via the integration of the IBM ILOG CPLEX Optimization Studio package, no-cost edition, and the high-performance MATLAB soft-ware.

5.3. Selected key performance indicators

Based on the data source mentioned before (ERSAR), we have considered the three criteria, seven sub-criteria, and twelve indicators listed in Table 5.1. In some cases, sub-criteria exhibit more than one indicator. Table 5.2 displays the nature of these twelve indicators: four desirable, four neutral, and four undesirable indicators. According to a regulator's recent technical report (2019), the same table identifies the optimal range associated with the best practices within the field.

Table 5.2
Nature and optimal ranges per indicator.
Source: Adapted from the technical report of ERSAR, January 2019.

Indicator	Nature[a]	Optimal range		
		Urban areas (8.7%)	Neither urban nor rural areas (26.1%)	Rural areas (65.2%)
RU01b	D	>95	>90	>80
RU02ab	D	>80	>70	>60
RU03b	U	<0.5	<0.5	<0.5
RU04b	N	6–24	6–24	6–24
RU06ab	N	100–110	100–110	100–110
RU07b	D	>100	>100	> 100
RU11b	U	<250 k	<250 k	<250 k
RU12b	N	400–500	400–500	400–500
RU13b[b]	N	1.0–2.5	1.0–3.0	1.0–3.5
RU14b	U	<4.5	<5.5	<6.5
RU16b	U	<13	<14	<15

[a] D, desirable; N, neutral; U, undesirable.
[b] A recent technical report published by ERSAR (January 2019) defines optimal ranges for RU13b depending upon two main aspects: rural vs. urban and selective collection vs. without the selective collection of urban waste. In our database, it is impossible to cluster utilities according to the latter issue. To be flexible, the optimal range is the reunion of ranges defined by the regulator to the utilities with and to the utilities without a selective collection of urban waste. For instance, ERSAR defines the optimal range for utilities with (resp. without) selective collection for urban areas as 1.5–2.5 (resp. 1.0–2.0); thus, we define the range as 1.0–2.5.

6. Results and discussion

6.1. The multiplicative BoD with regulatory constraints

6.1.1. The impact of simulation runs

As for any simulation, the number of runs to be performed in Algorithm 1 may impact on the final estimation of performance. Therefore, we conducted several experiments to measure the evolution of the average performance score of MSW management utilities as function of T, the number of Monte-Carlo cycles, for $T = 10^2, 5\times10^2, 10^3, 2\times10^3, 5\times10^3, 10^4, 2.5\times10^4, 5\times10^4, 10^5$. On the one hand, few iterations could not be sufficient to characterize the reality (in this case, the real values of observations). On the other hand, too many iterations are computationally expensive. After a certain number of runs, increasing it is usually irrelevant because the performance levels tend to stabilize. Therefore, we should seek the minimum number of iterations after which no substantial changes are observed: $\Delta_j^{T,T'} \leqslant 1\%$ for any j and $T' > T$. Table 6.1 and Fig. 6.1 portray the evolution of performance scores of six utilities, and we can observe that from $T = 5000$ iterations onward there is no change on estimated (average) performance. In some cases, the stabilization occurs for smaller values of T. The same was observed for the remaining 226 entities. Consequently, there is no gain in increasing T beyond that threshold. Note that it is well-known, in benchmarking literature, that a single observation can influence the performance of others. Therefore, it is straightforward to conclude that one observation featured by the imperfect knowledge of data is sufficient to cause performance variability in the remaining ones, except (of course) the benchmarks. These instead neither depend on the data quality nor the relative position of the poor performance entities. Given these results, we consider $T = 5000$ in the next subsections.

6.1.2. Performance levels

Table 6.2 provides the main statistics of performance scores of 11 selected utilities, including the average, standard deviation, and the factor $\mathrm{F}(\theta^j \geqslant 0.95)$, *i.e.*, the likelihood of an utility reach the performance score of 0.99 (or 99%). Performance scores were obtained through the multiplicative BoD model with regulatory constraints. Three (out of 11) utilities are classified as benchmarks, as they always exhibited an unitary performance score, regardless of the other observations. Two are potential benchmarks, with probabilities of 74 and 95% of being close to or in the frontier. The other six exhibit room for performance improvement, as their maximum scores in five thousand iterations were far from the 0.99-level.

Table 6.3 shows some basic statistics on performance scores obtained through the multiplicative BoD with regulatory constraints. The results were clustered with respect to the intervention area typology, as in Table 5.2. Fig. 6.2 shows the distribution of utilities performance classification per typology. We can see that there are benchmarks in each typology, *i.e.*, it is possible to find best practices regardless of the context in which utilities operate. Nonetheless, utilities located in urban areas exhibit larger performance scores, being followed by the entities operating in moderately urbanized areas. Although the variability of scores is substantial within the three groups (with special emphasis for the rural cases), the Kruskal–Wallis statistical test returned a p-value of 3.5% for a null hypothesis of equal distributions of performance scores among groups. It means, for a significance level of 5%, that indeed utilities in urban areas seem to perform better than their counterparts located in the mainland. These results are in line with the ones of Amaral et al. (2022), who attribute the better results of utilities in urban areas to the exploitation of economies of scale and density and the reduction of unitary costs and a better management of fleets. However, their results only considered efficiency, while our case study covers a broader range of performance dimensions.

Perhaps more important than estimating performance scores is the capacity of deriving targets for each KPI used in the model. Table 6.4 displays the basic statistics over the optimal targets obtained with the multiplicative BoD with regulatory constraints. It is noteworthy that all targets belong to the intervals defined in Table 5.2, independently from the intervention area typology. It means that our model constructs a frontier containing the "optimal" (mostly fictitious) utilities and obeying to all regulatory constraints. A comparison between the current average values and the optimal targets allows us to determine the relative room for improvement, as detailed in Table 6.5. Green (red) entries identify the required improvement (deterioration) in KPIs. Therefore, in overall terms, utilities need to:

1. Increase by 10.2% the physical accessibility of the service, *i.e.*, the rate of households served with refuse waste collection service, which can be done by increasing the number of disposal sites, reducing the distance between consecutive sites;
2. Increase by 49.2% the accessibility of the selective collection service, by largely increasing the number of selective waste disposal sites;
3. Increase by 68.6% the average burden associated with the municipal waste management service in terms of the household disposable income; this is, unquestionably, an undesirable KPI

Table 6.1
Performance evolution with T for six selected utilities.

T	j = 15	$\Delta_j^{T,T'}$	j = 34	$\Delta_j^{T,T'}$	j = 127	$\Delta_j^{T,T'}$	j = 156	$\Delta_j^{T,T'}$	j = 201	$\Delta_j^{T,T'}$	j = 230	$\Delta_j^{T,T'}$
100	0.652		0.741		0.917		0.844		0.538		0.209	
500	0.838	29%	0.758	2%	0.921	0%	0.856	1%	0.771	43%	0.535	156%
1000	0.748	−11%	0.580	−23%	0.923	0%	0.940	10%	0.781	1%	0.728	36%
2000	0.752	1%	0.860	48%	0.923	0%	0.867	−8%	0.781	0%	0.728	0%
5000	0.794	6%	0.860	0%	0.923	0%	0.867	0%	0.781	0%	0.728	0%
10,000	0.794	0%	0.860	0%	0.923	0%	0.867	0%	0.781	0%	0.728	0%
25,000	0.794	0%	0.860	0%	0.923	0%	0.867	0%	0.781	0%	0.728	0%
50,000	0.794	0%	0.860	0%	0.923	0%	0.867	0%	0.781	0%	0.728	0%
100,000	0.794	0%	0.860	0%	0.923	0%	0.867	0%	0.781	0%	0.728	0%

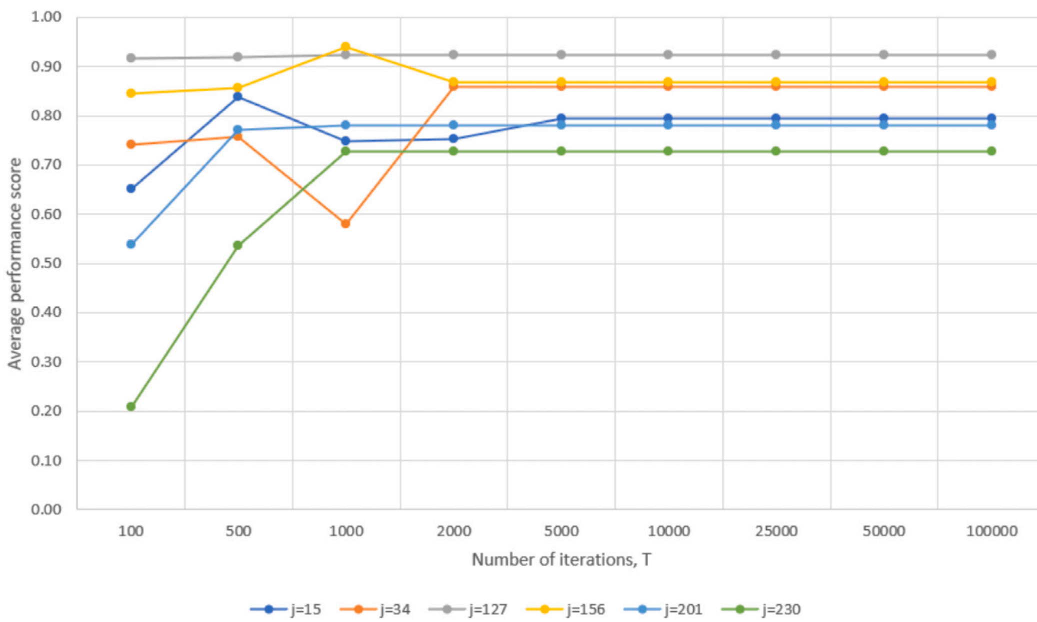

Fig. 6.1. Performance evolution with T for six selected utilities.

Table 6.2
Main statistics of performance scores of 11 selected utilities.

	Average	Std.Dev.	CV	min	max	F($\theta^j \geqslant 0.99$)	Classification
j = 15	0.794	0.154	19%	0.005	0.947	0%	Open to improvement
j = 21	1.000	0.000	0%	1.000	1.000	100%	Benchmark
j = 34	0.860	0.109	13%	0.410	0.877	0%	Open to improvement
j = 57	1.000	0.000	0%	1.000	1.000	100%	Benchmark
j = 127	0.923	0.139	15%	0.012	0.940	0%	Open to improvement
j = 128	0.944	0.089	9%	0.481	1.000	95%	Potential benchmark
j = 148	0.867	0.109	13%	0.283	1.000	74%	Potential benchmark
j = 156	0.867	0.129	15%	0.330	0.901	0%	Open to improvement
j = 201	0.780	0.021	3%	0.711	0.869	0%	Open to improvement
j = 215	1.000	0.000	0%	1.000	1.000	100%	Benchmark
j = 230	0.728	0.063	9%	0.522	0.944	0%	Open to improvement

whose increasing should be carefully studied — however, as long as the suggested burden raise obeys the regulator's constraint, *i.e.*, below 0.5, utilities should be free to implement this change, which, in turn, should be used to improve their performance in other dimensions, like the accessibility of the service and the vehicle renewal;

4. Increase the quality of the service, namely the refuse waste containers cleaning by 102.1%, by doubling the number of washes per deposit, both surface and underground;
5. Increase the coverage of expenses (associated with the increase of the average burden) via income (tariffs and subsidies) by 51.9%;
6. Increase recycling by 17.4%, which can result from augmenting the number of sites for sorted waste disposal and collection, *i.e.*, the accessibility of the selective collection service;
7. Decrease the cumulative miles of waste collection per vehicle by 28% before substitution, *i.e.*, vehicles need to be replaced sooner;

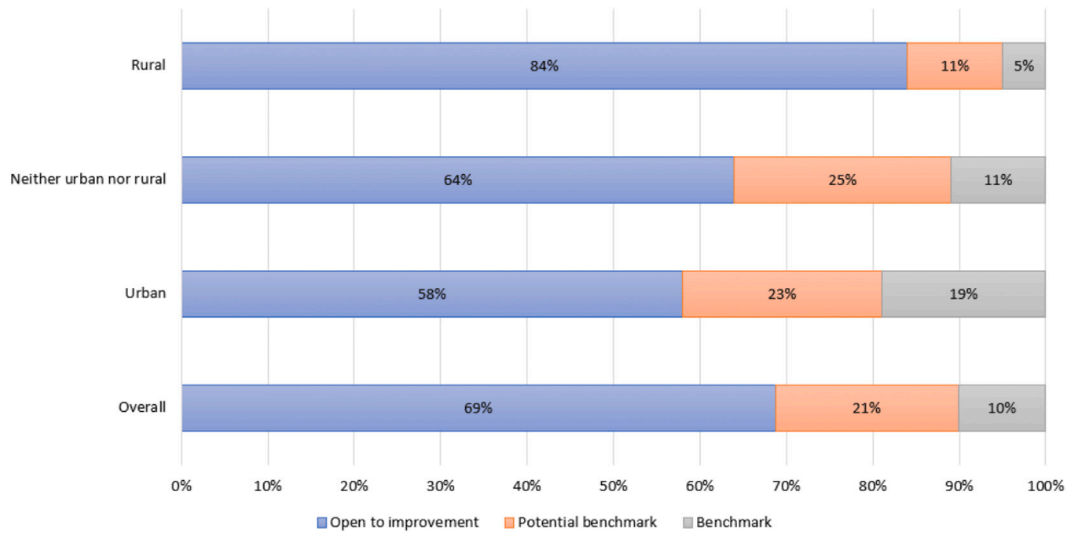

Fig. 6.2. Distribution of utilities classification per intervention area typology, following the multiplicative BoD model with regulatory constraints.

Table 6.3
Basic statistics on performance scores obtained through the multiplicative BoD with regulatory constraints.

	Urban	Neither urban nor rural	Rural	Overall
Average performance	0.917	0.889	0.714	0.846
Std. Deviation	0.388	0.319	0.405	0.347
CV	42%	36%	57%	41%
Min	0.601	0.599	0.424	0.553
Max	1.000	1.000	1.000	1.000

8. Increase the monetization of vehicles by 5.7%, by ensuring that vehicles are full or nearly full before returning the facilities, which may require route optimization (Lavigne et al., 2021) – the monetization may be improved through the physical accessibility of the service, either for refuse or for selective waste;
9. Decrease of staff allocated to waste management by 54.6%, which can result from increasing collected waste or simply by reducing the number of employees and outsourced staff;
10. Decrease the use of energy resources (fuel) by 14.4%, opting by electric vehicles, for instance;
11. Decrease the emissions of greenhouse gases from vehicles by 29.9%, which is in line with the previous topic.

6.2. *The multiplicative BoD without regulatory constraints*

If we relax the regulatory constraints and use the simple multiplicative BoD model, we observe an increase in performance scores, with reduction of variability; see Table 6.6. Fig. 6.3 shows an increase in the number of benchmarks and potential benchmarks, when compared to the model with regulatory constraints (see Fig. 6.2). This is a natural conclusion as the model becomes less constraint. However, we verified that about half of optimal targets violate the impositions made by the regulator, which implies that the performance scores are not valid as they are based on invalid constructs (biased frontier), not responding to the regulator's needs.

Table 6.4
Basic statistics over the optimal targets obtained with the multiplicative BoD with regulatory constraints.

	RU*										
	01b	02ab	03b	04b	06ab	07b	11b	12b	13b	14b	16b
Urban areas											
Average	96.5	83.8	0.3	12.9	100.9	110.1	238k	438.4	1.7	4.4	11.5
Min	95.0	80.0	0.2	6.0	100.0	100.0	201k	400.0	1.0	3.7	11.2
Max	99.4	95.7	0.5	24.0	110.0	115.1	250k	500.0	2.5	4.5	13.0
Neither urban nor rural areas (moderately urbanized areas)											
Average	94.0	74.0	0.3	6.8	106.1	101.5	241k	421.5	1.9	5.0	13.3
Min	90.0	70.0	0.2	6.0	100.0	100.0	220k	400.0	1.0	5.0	12.5
Max	96.7	80.2	0.5	24.0	110.0	117.9	250k	500.0	3.0	5.5	14.0
Rural areas											
Average	91.4	78.2	0.2	10.4	108.5	101.5	247k	459.5	2.8	6.2	13.3
Min	80.0	60.0	0.2	6.0	100.0	100.0	209k	400.0	1.0	5.9	13.2
Max	91.5	95.7	0.5	24.0	110.0	115.4	250k	500.0	3.5	6.5	15.0
Overall sample											
Average	93.5	76.0	0.3	8.3	106.3	102.2	242k	432.9	2.1	5.3	13.1
Min	80.0	60.0	0.2	6.0	100.0	100.0	201k	400.0	1.0	3.7	11.2
Max	99.4	95.7	0.5	24.0	110.0	117.9	250k	500.0	3.5	6.5	15.0

7. Concluding remarks

This paper proposed a multiplicative BoD model to construct composite indicators while accounting for regulatory constraints, undesirable KPIs, and imperfect knowledge of data. Perhaps the most important development here was the introduction of regulation-required constraints that considerably impact on entities performance assessment. Until now, research has disregarded this topic. Differently from the existing models, entities must obey the regulatory constraints in order to become efficient — it is a *sine qua non* condition for being classified as a benchmark. Our experiments have shown that, in such a case, many entities could fail to convey the requirements made by regulatory bodies. It implies that the frontier of benchmarks may have been misconstructed, and the performance scores biased. Therefore, in practice, the performance scores are lower because the model is more restrictive, but this fact can be beneficial for searching and adopting best practices and even for optimizing paid tariffs. The biasing

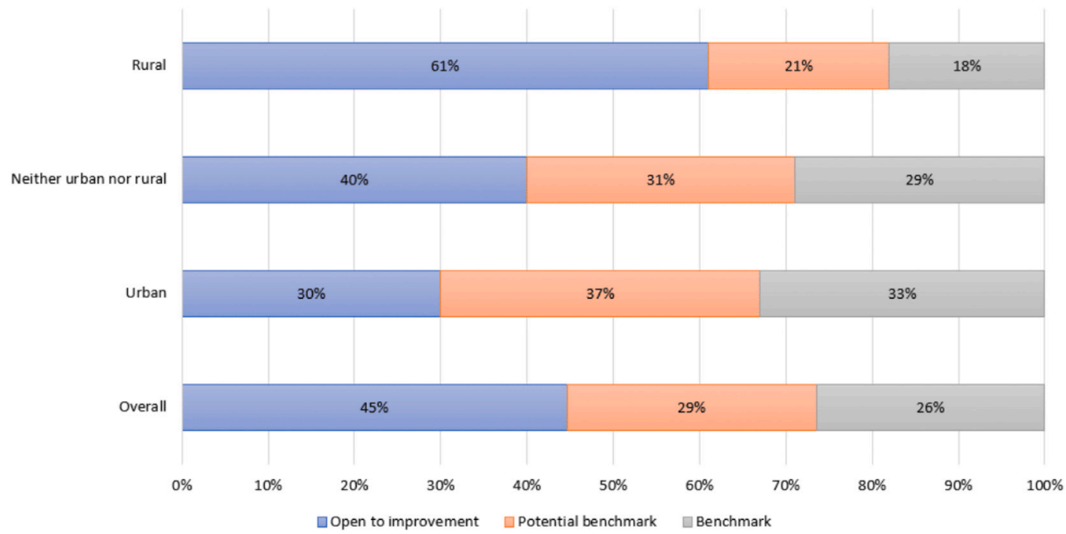

Fig. 6.3. Distribution of utilities classification per intervention area typology, following the multiplicative BoD model without regulatory constraints.

Table 6.5
Average improvement from current values to optimal targets (based on averages).

RU*	Average values				Average improvement			
	Overall	Urban	NUNR	Rural	Overall	Urban	NUNR	Rural
01b	84.86	85.03	84.80	84.95	10.2%	13.5%	10.8%	7.6%
02ab	50.93	50.87	50.78	50.90	49.2%	64.6%	45.8%	53.6%
03b	0.18	0.18	0.18	0.18	68.6%	91.3%	80.1%	33.5%
04b	4.08	4.07	4.04	4.05	102.1%	217.4%	67.5%	156.8%
06ab	69.97	69.40	69.50	69.72	51.9%	45.4%	52.6%	55.7%
07b	87.08	86.62	86.96	86.89	17.4%	27.1%	16.7%	16.8%
11b	336k	336k	338k	337k	−28.0%	−29.2%	−28.7%	−26.7%
12b	409.55	409.36	409.11	408.55	5.7%	7.1%	3.0%	12.5%
13b	4.63	4.67	4.67	4.66	−54.6%	−62.9%	−60.1%	−39.2%
14b	6.17	6.20	6.20	6.20	−14.4%	−28.3%	−18.9%	0%
16b	18.70	18.77	18.76	18.72	−29.9%	−38.6%	−29.3%	−29.2%

Note: green entries identify improvement, while red entries identify deterioration.

Table 6.6
Basic statistics on performance scores obtained through the multiplicative BoD without regulatory constraints.

	Urban	Neither urban nor rural	Rural	Overall
Average performance	0.980	0.941	0.899	0.933
Std. Deviation	0.174	0.113	0.206	0.199
CV	18%	12%	23%	21%
Min	0.877	0.901	0.884	0.894
Max	1.000	1.000	1.000	1.000

degree may of course depend on how restrictive the regulator is. In opposition, our framework constructs a frontier such that all targets with respect to it obey all constraints. Obviously, should the regulator be too restrictive, the model can result infeasible for some entities. Nonetheless, this is an issue to be solved in future research, although not observed in our empirical experiment.

Also, the new model included a framework useful to deal with the common problem of imperfect knowledge of data, which hampers most performance assessment models and casts doubts about the validity of the existing results (simply because the quality of data conditions the quality of the output). However, by using the basic logic of Monte-Carlo simulation and a simplistic framework that anyone can implement in any programming language, the new model can easily address this data-related problem. Actually, this framework can be straightforwardly adapted to any performance analysis featured by some imperfect knowledge of data, which makes its acceptance easy.

Besides the theoretical development carried out in this research, the paper also contributed to the literature with the empirical study of the Portuguese urban solid waste management utilities. Important results were derived, namely the need to improve the physical accessibility of the service via introducing more waste disposal sites, the vehicles routing and renewal that should occur at smaller cumulative distances, the need to reduce greenhouse emissions and fuel utilization, and importantly, the requirement to augment the average burden over household disposable income, which should be used to improve the service's overall quality. In practical terms, we present five main strategic directions that utility management bodies can follow to improve performance levels: (1) searching for routes optimization to save fuel and minimize pollution; (2) purchasing new or rebranded electric vehicles for waste collection, replacing many of the existing fleets as soon as possible, to avoid excessive cumulative miles and to mitigate the considerable current greenhouse gas emissions resulting from the collecting operation; (3) increasing the number of disposal sites, either for refuse or selective waste, by decreasing the distance between two consecutive sites, while launching marketing strategies to promote recycling; (4) cleaning more often the existing disposal deposits to prevent stink, household dissatisfaction, damages caused by stray animals, and even zoonotic diseases (public health issues); and (5) optimizing payments due to the service, aligning the performance levels with the regulator's objectives. Regarding this last strategic direction, it is paramount to consider the operator's technical efficiency and effectiveness during the optimal tariff computation such that consumers (households) are not unfairly charged for the inefficient use of resources. As long as the regulation-based model proposed in this study is adopted with the goal of optimizing those payments, it becomes clear that the identified inefficiency levels are higher than those obtained with another model disregarding those constraints. Therefore, the proposed model clearly contributes to social welfare, making the due payments fairer to citizens. However, the empirical application has shown that the average burden associated with the municipal management service must increase. Indeed, the need to optimize tariffs does not imply its reduction; rather, it implies that one should pay only for an efficient and effective service.

CRediT authorship contribution statement

Diogo Cunha Ferreira: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Paulo Caldas:** Formal analysis, Funding acquisition, Investigation, Methodology, Resources, Validation, Visualization, Writing – original draft, Writing – review & editing. **Miguel Varela:** Funding acquisition, Investigation, Resources, Supervision, Validation, Visualization, Writing – review & editing. **Rui Cunha Marques:** Data curation, Investigation, Project administration, Supervision, Validation, Visualization, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data used in the case study are available at: https://www.ersar.pt/pt/publicacoes/relatorio-anual-do-setor [In Portuguese].

Acknowledgment

Any errors are the authors' responsibility. The usual disclaimer applies.

References

Agasisti, T., Barra, C., & Zotti, R. (2019). Research, knowledge transfer, and innovation: The effect of Italian universities' efficiency on local economic development 2006–2012. *Journal of Regional Science*, *59*(5), 819–849.

Amado, C. A., São José, J. M., & Santos, S. P. (2016). Measuring active ageing: A data envelopment analysis approach. *European Journal of Operational Research*, *255*(1), 207–223.

Amaral, C., Isabel Pedro, M., Cunha Ferreira, D., & Cunha Marques, R. (2022). Performance and its determinants in the Portuguese municipal solid waste utilities. *Waste Management*, *139*, 70–84.

Banker, R., & Maindiratta, A. (1986). Piecewise loglinear estimation of efficient production surfaces. *Management Science*, *32*, 126–135.

Bernini, C., Guizzardi, A., & Angelini, G. (2012). DEA-like model and common weights approach for the construction of a subjective community well-being indicator. *Social Indicators Research*, *114*, 405–424.

Bernini, C., Guizzardi, A., & Angelini, G. (2013). DEA-like model and common weights approach for the construction of a subjective community well-being indicator. *Social Indicators Research*, *114*, 405–424.

Blancard, S., & Hoarau, J. F. (2013). A new sustainable human development indicator for small island developing states: A reappraisal from data envelopment analysis. *Economic Modelling*, *30*, 623–635.

Blancas, F., Contreras, I., & Ramirez-Hurtado, J. (2013). Constructing a composite indicator with multiplicative aggregation under the objective of ranking alternatives. *Journal of the Operational Research Society*, *64*, 668–678.

Bogetoft, P., & Otto, L. (2011). *Benchmarking with data envelopment analysis, stochastic frontier analysis and R*. New York: International series in operations research & management science, Springer.

Buchanan, J. T., Henig, E. J., & Henig, M. I. (1998). Objectivity and subjectivity in the decision making process. *Annals of Operations Research*, *80*, 333–345.

Bullen, P. (2003). *Handbook of means and their inequalities*. Netherlands: Kluwer.

Cai, M., Guy, C., Héroux, M., Lichtfouse, E., & An, C. (2021). The impact of successive COVID-19 lockdowns on people mobility, lockdown efficiency, and municipal solid waste. *Environmental Chemistry Letters*.

Caldas, P., Ferreira, D., Dollery, B., & Marques, R. C. (2019). Are there scale economies in urban waste and wastewater municipal services? A non-radial input-oriented model applied to the Portuguese local government. *Journal of Cleaner Production*, *2019*, 531–539.

Carvalho, P., & Marques, R. C. (2014). Economies of size and density in recycling of municipal solid waste in Portugal. *Waste Management*, *34*, 12–20.

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

Chambers, R., Chung, Y., & Färe, R. (1998). Profit, directional distance functions, and Nerlovian efficiency. *Journal of Optimization Theory and Applications*, *98*, 351–364.

Charles, V., Aparicio, J., & Zhu, J. (2019). The curse of dimensionality of decision-making units: A simple approach to increase the discriminatory power of data envelopment analysis. *European Journal of Operational Research*, *279*(3), 929–940.

Cherchye, L., Moesen, W., Rogge, N., & Puyenbroeck, T. V. (2007). An introduction to 'benefit of the doubt' composite indicators. *Social Indicators Research*, *82*, 111–145.

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

Coco, G., Lagravinese, R., & Resce, G. (2020). Beyond the weights: A multicriteria approach to evaluate inequality in education. *The Journal of Economic Inequality*, *18*, 469–489.

Daraio, C., & Simar, L. (2014). Directional distances and their robust versions: Computational and testing issues. *European Journal of Operational Research*, *237*(1), 358–369.

de Castro-Pardo, M., Martínez, P. F., & Zabaleta, A. P. (2022). An initial assessment of water security in Europe using a DEA approach. *Sustainable Technology and Entrepreneurship*, *1*(1), Article 100002.

Debreu, G. (1960). Topological methods in cardinal utility theory. In K. Arrow, S. Karlin, & P. Suppes (Eds.), *Mathematical methods in social sciences*. Stanford: Stanford University Press.

Decancq, K., & Lugo, M. (2013). Weights in multidimensional indices of wellbeing: An overview. *Econometric Reviews*, *32*, 7–34.

Delgado-Antequera, L., Gémar, G., Molinos-Senante, M., Gómez, T., Caballero, R., & Sala-Garrido, R. (2021). Eco-efficiency assessment of municipal solid waste services: Influence of exogenous variables. *Waste Management*, *130*, 136–146.

Despotis, D. (2002). Improving the discriminating power of DEA: Focus on globally efficient units. *Journal of Operations Research Society*, *53*, 314–323.

Despotis, D., & Smirlis, Y. (2002). Data envelopment analysis with imprecise data. *European Journal of Operational Research*, *140*(1), 24–36.

D'Inverno, G., Carosi, L., & Romano, G. (2021). Environmental sustainability and service quality beyond economic and financial indicators: A performance evaluation of Italian water utilities. *Socio-Economic Planning Sciences*, *75*, Article 100852.

D'Inverno, G., & De Witte, K. (2020). Service level provision in municipalities: A flexible directional distance composite indicator. *European Journal of Operational Research*, *286*(3), 1129–1141.

Dominguez-Gil, C., Segovia-Gonzalez, M. M., & Contreras, I. (2021). A multiplicative composite indicator to evaluate educational systems in OECD countries. *Compare: A Journal of Comparative and International Education*, 1–18.

Du, J., Wang, J., Chen, Y., Chou, S. Y., & Zhu, J. (2014). Incorporating health outcomes in pennsylvania hospital efficiency: An additive super-efficiency DEA approach. *Annals of Operational Research*, *221*, 161–172.

Elster, C., & Klauenberg, K. (2020). A quantile inequality for location-scale distributions. *Statistics & Probability Letters*, *165*, Article 108851.

Emrouznejad, A., & Cabanda, E. (2010). An aggregate measure of financial ratios using a multiplicative DEA model. *International Journal of Financial Services Management*, *4*, 114–126.

Färe, R., & Karagiannis, G. (2014). A postscript on aggregate Farrell efficiencies. *European Journal of Operational Research*, *233*(3), 784–786.

Färe, R., Karagiannis, G., Hasannasab, M., & Margaritis, D. (2019). A benefit-of-the-doubt model with reverse indicators. *European Journal of Operational Research*, *278*(2), 394–400.

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

Ferreira, D. C., Figueira, J. R., Greco, S., & Marques, R. (2021). Data envelopment analysis models with imperfect knowledge of input and output values: An application to Portuguese public hospitals. arXiv preprint arXiv:2104.09890.

Ferreira, D. C., Graziele, I., Marques, R. C., & Gonçalves, J. (2021). Investment in drinking water and sanitation infrastructure and its impact on waterborne diseases dissemination: The Brazilian case. *Science of the Total Environment*, *779*, Article 146279.

Ferreira, D., & Marques, R. (2021). Public-private partnerships in health care services: Do they outperform public hospitals regarding quality and access? Evidence from Portugal. *Socio-Economic Planning Sciences*, *73*, Article 100798.

Ferreira, D., Marques, R., & Nunes, A. (2018). Economies of scope in the health sector: The case of Portuguese hospitals. *European Journal of Operational Research*, *266*(2), 716–735.

Ferreira, D., Marques, R., & Nunes, A. (2021). Pay for performance in health care: A new best practice tariff-based tool using a log-linear piecewise frontier function and a dual–primal approach for unique solutions. *Operational Research*, *21*, 2101–2146.

Ferreira, D. C., Marques, R. C., & Pedro, M. I. (2018). Explanatory variables driving the technical efficiency of European seaports: An order-α approach dealing with imperfect knowledge. *Transportation Research Part E: Logistics and Transportation Review*, *119*, 41–62.

Ferreira, D. C., Marques, R. C., Pedro, M. I., & Amaral, C. (2020). Economic inefficiency levels of urban solid waste management services in Portugal. *Sustainability*, *12*(10), 4170.

Ferreira, D., Nunes, A., & Marques, R. (2021). Optimising payments based on efficiency, quality, complexity, and heterogeneity: The case of hospitals funding. *International Transactions of Operation Research*, *27*, 1930–1961.

French, S. (1995). Uncertainty and imprecision: Modelling and analysis. *Journal of the Operational Research Society*, *46*(1), 70–79.

Fusco, E. (2015). Enhancing non-compensatory composite indicators: A directional proposal. *European Journal of Operational Research*, *242*(2), 620–630.

Fusco, E., Vidoli, F., & Rogge, N. (2020). Spatial directional robust benefit of the doubt approach in presence of undesirable output: An application to Italian waste sector. *Omega*, *94*, Article 102053.

Fusco, E., Vidoli, F., & Sahoo, B. K. (2018). Spatial heterogeneity in composite indicator: A methodological proposal. *Omega*, *77*, 1–14.

Giambona, F., & Vassallo, E. (2014). Composite indicator of social inclusion for European countries. *Social Indicators Research*, *116*, 269–293.

Greco, S., Ehrgott, M., & Figueira, J. R. (2016). *Multiple criteria decision analysis: State of the art surveys*. Springer-Verlag New York.

Greco, S., Ishizaka, A., Tasiou, M., & Torrisi, G. (2019). On the methodological framework of composite indices: A review of the issues of weighting, aggregation, and robustness. *Social Indicators Research*, *141*, 61–94.

Guardiola, J., & Picazo-Tadeo, A. J. (2014). Building weighted-domain composite indices of life satisfaction with data envelopment analysis. *Social Indicators Research*, *117*, 257–274.

Gulati, R., Kattumuri, R., & Kumar, S. (2020). A non-parametric index of corporate governance in the banking industry: An application to Indian data. *Socio-Economic Planning Sciences*, *70*, Article 100702.

Hardy, G., Littlewood, J., & Pólya, G. (1952). *Inequalities*. Cambridge, England: Cambridge University Press.

Hayek, F. A. (1945). The use of knowledge in society. *The American Economic Review*, *35*(4), 519–530.

Henriques, A. A., Camanho, A. S., Amorim, P., & Silva, J. G. (2020). Performance benchmarking using composite indicators to support regulation of the Portuguese wastewater sector. *Utilities Policy*, *66*, Article 101082.

Kao, C. (2017). *Network data envelopment analysis: Foundations and extensions*. Springer International Publishing.

Keeney, R., & Raiffa, H. (1993). *Decisions with multiple objectives: Preferences and value trade-offs*. Wiley series in probability and mathematical statistics. Applied probability and statistics, Cambridge University Press.

Knox Lovell, C., Pastor, J. T., & Turner, J. A. (1995). Measuring macroeconomic performance in the OECD: A comparison of European and non-European countries. *European Journal of Operational Research*, *87*(3), 507–518.

Krantz, D., Luce, R., Suppes, P., & Tversky, A. (1971). *Additive and polynomial representations*: *vol. 1*, *Foundations of measurement*. New York: Academic Press.

Kuosmanen, T. (2005). Weak disposability in nonparametric productivity analysis with undesirable outputs. *American Journal of Agricultural Economics*, *87*, 1077–1082.

Kuosmanen, T., & Podinovski, V. (2009). Weak disposability in nonparametric production analysis: Reply to färe and grosskopf. *American Journal of Agricultural Economics*, *91*, 539–545.

Lafuente, E., Alonso-Ubieta, S., Leiva, J., & Mora-Esquivel, R. (2021). Strategic priorities and competitiveness of businesses operating in different entrepreneurial ecosystems: A benefit of the doubt (BOD) analysis. *International Journal of Entrepreneurial Behavior & Research*, *27*(5), 1351–1377.

Lafuente, E., Araya, M., & Leiva, J. C. (2020). Assessment of local competitiveness: A composite indicator analysis of Costa Rican counties using the 'benefit of the doubt' model. *Socio-Economic Planning Sciences*, Article 100864.

Lavigne, C., Beliën, J., & Dewil, R. (2021). An exact routing optimization model for bio-waste collection in the Brussels capital region. *Expert Systems with Applications*, *183*, Article 115392.

Liu, J., Fang, S. C., & Chen, H. (2020). Multiplicative data envelopment analysis cross-efficiency and stochastic weight space acceptability analysis for group decision making with interval multiplicative preference relations. *Information Sciences*, *514*, 319–332.

Llanquileo-Melgarejo, P., & Molinos-Senante, M. (2021). Evaluation of economies of scale in eco-efficiency of municipal waste management: An empirical approach for Chile. *Environmental Science and Pollution Research*, *28*, 28337–28348.

Maria, C., Góis, J., & Leitão, A. (2020). Challenges and perspectives of greenhouse gases emissions from municipal solid waste management in angola. *Energy Reports*, *6*, 364–369.

Marques, R. C., & Simões, P. (2009). Incentive regulation and performance measurement of the Portuguese solid waste management services. *Waste Management & Research*, *27*, 188–196.

Marques, R. C., Simões, P., & Pinto, F. (2018). Tariff regulation in the waste sector: An unavoidable future. *Waste Management*, *78*, 292–300.

Matos, R., Ferreira, D., & Pedro, M. (2021). Economic analysis of Portuguese public hospitals through the construction of quality, efficiency, access, and financial related composite indicators. *Social Indicators Research*, *157*, 361–392.

Mauricio-Iglesias, M., Longo, S., & Hospido, A. (2020). Designing a robust index for WWTP energy efficiency: The enerwater water treatment energy index. *Science of the Total Environment*, *713*, Article 136642.

Mehdiloozad, M., Sahoo, B. K., & Roshdi, I. (2014). A generalized multiplicative directional distance function for efficiency measurement in DEA. *European Journal of Operational Research*, *232*(3), 679–688.

Melyn, W., & Moesen, W. (1991). *Towards a synthetic indicator of macroeconomic performance: Unequal weighting when limited information is available*: *Public economic working paper 17*, Belgium: Katholieke Universiteit Leuven.

Mergoni, A., D'Inverno, G., & Carosi, L. (2021). A composite indicator for measuring the environmental performance of water, wastewater, and solid waste utilities. *Utilities Policy*, Article 101285.

Moers, F. (2005). Discretion and bias in performance evaluation: The impact of diversity and subjectivity. *Accounting, Organizations and Society*, *30*(1), 67–80.

Mora, C., Manzini, R., Gamberi, M., & Cascini, A. (2014). Environmental and economic assessment for the optimal configuration of a sustainable solid waste collection system: A 'kerbside' case study. *Production Planning and Control*, *25*, 737–761.

Olay-Romero, E., Turcott-Cervantes, D. E., del Consuelo Hernández-Berriel, M., Lobo-García de Cortázar, A., Cuartas-Hernández, M., & de la Rosa-Gómez, I. (2020). Technical indicators to improve municipal solid waste management in developing countries: A case in Mexico. *Waste Management*, *107*, 201–210.

Olinsky, A., Chen, S., & Harlow, L. (2003). The comparative efficacy of imputation methods for missing data in structural equation modeling. *European Journal of Operational Research*, *151*(1), 53–79.

Peffers, K., Tuunanen, T., Rothenberger, M., & Chatterjee, S. (2008). A design science research methodology for information systems research. *Journal of Management Information Systems*, *24*(3), 45–77.

Pereira, M. A., Camanho, A. S., Figueira, J. R., & Marques, R. C. (2021). Incorporating preference information in a range directional composite indicator: The case of Portuguese public hospitals. *European Journal of Operational Research*, *294*(2), 633–650.

Pereira, M. A., Camanho, A. S., Marques, R. C., & Figueira, J. R. (2021). The convergence of the world health organization member states regarding the United Nations' sustainable development goal 'good health and well-being'. *Omega*, *104*, Article 102495.

Podinovski, V. V., & Kuosmanen, T. (2011). Modelling weak disposability in data envelopment analysis under relaxed convexity assumptions. *European Journal of Operational Research*, *211*(3), 577–585.

Raaijmakers, Q. A. W. (1999). Effectiveness of different missing data treatments in surveys with likert-type data: Introducing the relative mean substitution approach. *Educational and Psychological Measurement*, *59*(5), 725–748.

Reilly, M. (1993). Data analysis using hot deck multiple imputation. *Journal of the Royal Statistical Society, Series D*, *42*(3), 307–313.

Ríos, A. M., & Picazo-Tadeo, A. J. (2021). Measuring environmental performance in the treatment of municipal solid waste: The case of the European union-28. *Ecological Indicators*, *123*, Article 107328.

Rockafellar, R. (1970). *Princeton mathematical series*: *vol. 28*, *Convex analysis*. Princeton, N.J.: Princeton University Press.

Rogge, N. (2018a). Composite indicators as generalized benefit-of-the-doubt weighted averages. *European Journal of Operational Research*, *267*(1), 381–392.

Rogge, N. (2018b). On aggregating benefit of the doubt composite indicators. *European Journal of Operational Research*, *264*(1), 364–369.

Rogge, N., & Van Nijverseel, I. (2019). Quality of life in the European union: A multidimensional analysis. *Social Indicators Research*, *141*, 765–789.

Roy, B., Figueira, J., & Almeida-Dias, J. (2014). Discriminating thresholds as a tool to cope with imperfect knowledge in multiple criteria decision aiding: Theoretical results and practical issues. *Omega: The International Journal of Management Science*, *43*, 9–20.

Sahoo, B. K., Singh, R., Mishra, B., & Sankaran, K. (2017). Research productivity in management schools of India during 1968–2015: A directional benefit-of-doubt model analysis. *Omega*, *66*, 118–139.

Sala-Garrido, R., Mocholí-Arce, M., & Molinos-Senante, M. (2021). Assessing the quality of service of water companies: A 'benefit of the doubt' composite indicator. *Social Indicators Research*, *155*, 371–387.

Sarra, A., & Nissi, E. (2013). A spatial composite indicator for human and ecosystem well-being in the Italian urban areas. *Social Indicators Research*, *148*, 353–377.

Seiford, L. M., & Zhu, J. (2002). Modeling undesirable factors in efficiency evaluation. *European Journal of Operational Research*, *142*(1), 16–20.

Shwartz, M., Burgess, J. F., & Zhu, J. (2016). A DEA based composite measure of quality and its associated data uncertainty interval for health care provider profiling and pay-for-performance. *European Journal of Operational Research*, *253*(2), 489–502.

Silva, M. C., Camanho, A. S., & Barbosa, F. (2020). Benchmarking of secondary schools based on students' results in higher education. *Omega*, *95*, Article 102119.

Simões, P., Carvalho, P., & Marques, R. C. (2013). The market structure of urban solid waste services. How different models lead to different results. *Local Government Studies*, *39*, 396–413.

Simões, P., De Witte, K., & Marques, R. C. (2010). Regulatory structures and operational environment in the Portuguese waste sector. *Waste Management*, *30*(6), 1130–1137.

Smirlis, Y., Maragos, E., & Despotis, D. (2006). Data envelopment analysis with missing values: An interval DEA approach. *Applied Mathematics and Computation*, *177*(1), 1–10.

Tiedemann, T., Francksen, T., & Latacz-Lohmann, U. (2011). Assessing the performance of German Bundesliga football players: A non-parametric metafrontier approach. *Central European Journal of Operational Research*, *19*, 571–587.

Tofallis, C. (2014). On constructing a composite indicator with multiplicative aggregation and the avoidance of zero weights in DEA. *Journal of the Operational Research Society*, *65*, 791–792.

Tsolas, I. E., Charles, V., & Gherman, T. (2020). Supporting better practice benchmarking: A DEA-ANN approach to bank branch performance assessment. *Expert Systems with Applications*, *160*, Article 113599.

Van Puyenbroeck, T. (2018). On the output orientation of the benefit-of-the-doubt-model. *Social Indicators Research*, *139*, 415–431.

Van Puyenbroeck, T., & Rogge, N. (2017). Geometric mean quantity index numbers with benefit-of-the-doubt weights. *European Journal of Operational Research*, *256*(3), 1004–1014.

Verbunt, P., & Rogge, N. (2018). Geometric composite indicators with compromise benefit-of-the-doubt weights. *European Journal of Operational Research*, *264*(1), 388–401.

Yadav, V., & Karmakar, S. (2020). Sustainable collection and transportation of municipal solid waste in urban centers. *Sustainable Cities and Society*, *53*, Article 101937.

Yao, L., Shuai, Y., & Chen, X. (2020). Regional water system vulnerability evaluation: A bi-level DEA with multi-followers approach. *Journal of Hydrology*, *589*, Article 125160.

Zanella, A., Camanho, A. S., & Dias, T. G. (2015). Undesirable outputs and weighting schemes in composite indicators based on data envelopment analysis. *European Journal of Operational Research*, *245*(2), 517–530.

Zhou, P., Ang, B., & Poh, K. (2007). A mathematical programming approach to constructing composite indicators. *Ecological Economics*, *62*(2), 291–297.

Zhou, P., Ang, B., & Zhou, D. (2010). Weighting and aggregation in composite indicator construction: A multiplicative optimization approach. *Social Indicators Research*, *96*, 169–181.

Zhou, P., Fan, L. W., & Zhou, D. Q. (2010). Data aggregation in constructing composite indicators: A perspective of information loss. *Expert Systems with Applications*, *37*(1), 360–365.