

An evaluation of cross-efficiency methods: With an application to warehouse performance



Bert M. Balk^a, M.B.M. (René) De Koster^a, Christian Kaps^b, José L. Zofío^{a,c,*}

^a Rotterdam School of Management, Erasmus University, Rotterdam, the Netherlands

^b The Wharton School, The University of Pennsylvania, Philadelphia, USA

^c Department of Economics, Universidad Autónoma de Madrid, Madrid, Spain

ARTICLE INFO

Article history:

Received 30 October 2020

Revised 20 February 2021

Accepted 1 April 2021

Available online 24 April 2021

Keywords:

DEA

Cross-efficiency

Rank-order

Warehouse

ABSTRACT

Cross-efficiency measurement is an extension of Data Envelopment Analysis that allows for tie-breaking ranking of the Decision Making Units (DMUs) using all the peer evaluations. In this article we examine the theory of cross-efficiency measurement by comparing a selection of methods popular in the literature. These methods are applied to performance measurement of European warehouses. We develop a cross-efficiency method based on a rank-order DEA model to accommodate the ordinal nature of some key variables characterizing warehouse performance. This is one of the first comparisons of methods on a real-life dataset and the first time that a model allowing for qualitative variables is included in such a comparison. Our results show that the choice of model matters, as one obtains statistically different rankings from each one of them. This holds in particular for the multiplicative and game-theoretic methods whose results diverge from the classic method. From a managerial perspective, focused on the applicability of the methods, we evaluate them through a multidimensional metric which considers their capability to rank DMUs, their ease of implementation, and their robustness to sensitivity analyses. We conclude that standard weight-restriction methods, as initiated by Sexton et al. [48], perform as well as recently introduced, more sophisticated alternatives.

© 2021 The Author(s). Published by Elsevier Inc.
This is an open access article under the CC BY license
(<http://creativecommons.org/licenses/by/4.0/>)

1. Introduction

Although Data Envelopment Analysis (DEA) is a powerful method to study productive performance, it has several shortcomings when it comes to decision making. One of the most important weaknesses from a managerial perspective is that the DEA method returns, as virtual benchmarks, processes that employ unrealistic quantities of inputs and outputs. This translates into rankings of performance that are meaningless if these underlying benchmarks were considered as valid references without further inspection. This is because, in its multiplier formulation, DEA searches for the most favorable weights (shadow prices) when evaluating a production unit, thereby frequently assigning zero values to certain variables when constructing the 'virtual' aggregate output to input productivity ratio – each constructed as a linear combination of observed

* Corresponding author at: Rotterdam School of Management, Erasmus University, Rotterdam, the Netherlands.
E-mail addresses: jose.zofio@uam.es, jzofio@rsm.nl (J.L. Zofío).

magnitudes.¹ While the original weight flexibility behind this 'self-appraisal' is one of the most attractive aspects of the method, it often leads to unreasonable results, namely if the optimal weights are not consistent with prior knowledge of the production process, or if some inputs or outputs are ignored in the analysis.

A second consequence of this weight flexibility is that, when searching for the optimal weights, a large number of production units are deemed efficient by default. Eventually, any unit using the smallest quantity of any input, or producing the largest quantity of any output is categorized as efficient, regardless of the use it makes of other inputs, and the level of production of other outputs (which are assigned zero weights). A similar reasoning goes for inefficient units whose efficiency is overstated because improbable weights are taken into account. This implies that the obtained ranking of production units misrepresents best-practice performance, as units whose 'virtual' production processes are either implausible from a managerial perspective, or infeasible from an engineering perspective (e.g., warehouse service production without floor space) may be signaled as efficient. Ultimately, the flexibility of DEA may turn against the method itself by hampering its discriminatory power, a problem that is aggravated when the degrees of freedom are limited, as a result of a small number of observations relative to the number of inputs and outputs. This makes it difficult to draw conclusions on best practices.

To remedy these shortcomings, several proposals have been developed to improve the discriminatory power of DEA. [2] provided a comprehensive review of the literature. Existing methods were classified in ten different categories. The first category corresponds to the cross-efficiency methods that are the subject of our study. The second category concerns the super-efficiency approach initiated by [4]. In this method, each efficient DMU in turn is removed from the reference set, thereby obtaining an efficiency score greater than one, which allows to break the tie among the scores. This method will also be used in our empirical section. The third category ranks DMUs according to their relative importance to inefficient units; for instance, how often they serve as reference peers for inefficient units (see [51]). The fourth category relies on statistical techniques (such as canonical correlation analysis) directly applied after running a DEA model; see [33]. The fifth category focuses on ranking inefficient DMUs rather than efficient ones. This is accomplished in terms of a so-called 'measure of efficiency dominance', based on the magnitudes of the slacks obtained from an additive model, as proposed by Bardhan et al. [13]. The sixth category resorts to multilevel DEA within multicriteria decision making methods, complemented with analytic hierarchy processes, as in [34]. The seventh category brings inefficient frontiers into the analysis by solving 'inverted' DEA models, which measure inefficiency with respect to reference hyperplanes defined by the worst performing DMUs, [59]. Within the so-called TOPSIS methods, the eighth category considers virtual DMUs identifying the best (ideal) and worst (anti-ideal) performance. The ranking in this case takes into consideration both how close and how far the DMUs are from these two benchmarks, respectively – see [12,17]. The ninth category contains methods where decision makers bring external judgments to the evaluation process in the form of weight restrictions. These weights, imposing bounds on the shadow prices of inputs and outputs, are included in the multiplier form of the DEA problem – for successive reviews see [3,23,49]. The final category is based on fuzzy methods, implying that the input and output multipliers are considered as fuzzy sets, as well as the final (aggregate) virtual input and output obtained in the standard DEA model, [6].

A drawback shared by most of these proposals is that the optimal weights, obtained by solving the DEA models, are not unique. The existence of an infinite number of solutions besides the one obtained by the simplex method in a given run creates uncertainty in the evaluation process, and may lead to conflicting prescriptions from a managerial perspective (e.g., in the form of multiple rates of substitution between inputs or transformation between outputs). Thus, it is relevant to have some criterion for selecting a specific set of weights among all optimal solutions. Ideally, such a criterion should also solve the ranking arbitrariness and provide a meaningful multilateral comparison of efficiencies among the units. In addition, as concluded in our methodological section, this comparison should be consistent with index number theory from the perspective of productivity measurement. Ultimately, as the final objective of DEA is to compare performance among production units, the more bilateral evaluations are brought into the analysis, the more robust the rankings are to partial (mis)representations of the production technology, as well as to extreme or unobserved production units (such as the ideal and anti-ideal units).

Cross-efficiency measurement is an approach based on multilateral comparison of efficiency, yielding a consistent ranking without truncated efficiency values, thereby improving discrimination, and based on a specific well-identified criterion with a meaningful productivity interpretation. Introduced by Sexton et al. [48] and popularized by Doyle and Green [26], cross-efficiency measurement chooses a specific set of weights, making use of a two-stage process. In the first 'self-appraisal' stage, for each production unit its standard efficiency score is computed. In the second stage, the weights are selected to globally maximize or minimize the efficiency scores of all the competitors in the industry (the so-called benevolent and aggressive approach, respectively), while keeping the efficiency score of the evaluated unit unchanged. The basic idea of cross-efficiency measurement is to compare each unit with all its rivals, using all their weights rather than only its own weights. Finally, the cross-efficiency score of the unit is calculated as the (geometric) mean of all its cross-efficiencies.

Recent surveys of these methods are those by Cook and Seiford [19], Cook and Zhu [22], Zhu [60], Liu et al. [41] identified cross-efficiency measurement as one of the four research fronts in DEA. The present study contributes in several methodological, computational, empirical, and managerial dimensions to theory and practice. From a *methodological perspective* we explore the relative merits of alternative cross-efficiency methods by focusing on three areas.

¹ Moreover, as remarked by [23], zero optimal weights in the multiplier formulation correspond, by duality, to non-zero slacks in the primal envelopment form of the DEA model, and therefore the evaluated unit is assessed with respect to a benchmark that does not belong to the Pareto-efficient frontier. On this we refer the reader to [54].

First, we compare the results obtained by five representative methods using real-life data. The first class of methods can be referred to as ‘weight selection models’ which includes the already mentioned classic approaches of [26,48]. The benevolent and aggressive secondary goals have been modified by other authors, including the ‘multiplicative DEA approach’ by Charnes et al. [15,16], Cook and Zhu [21], which constitute the second class of models. In the last decades [39,56] re-interpreted the second-stage solution from the perspective of non-cooperative and cooperative games, giving rise to the third class of ‘game-theoretic’ models. Wu et al. [57] extended this line of research by considering a bargaining game. Ramón et al. [47], Wang and Chin [52] determined the weights for each DMU without considering the impact on its rivals. The game-theoretic cross-efficiency method was further extended and applied by Ma et al. [43]. From each of the above classes we select one of the core, most representative models for implementation.

Second, since in many applications some of the input or output variables are qualitative, we link the ordinal model introduced by Cook and Zhu [20] to the cross-efficiency measurement literature. Third, by revisiting the definition of cross-efficiency, we reinterpret it as a measure comparing the relative productivity of any two observations, and advocate the use of the geometric mean as aggregator function for the cross-efficiency scores.

From a *computational perspective*, we notice that there does not exist a single set of functions implementing cross-efficiency models in a common environment. To fill this gap and provide researchers with a suitable toolbox, we have programmed the functions solving the various cross-efficiency models. The result is available as free software, under the GNU General Public License version 3, which can be downloaded from <https://github.com/joselzofio/DEACrossEfficiencyMATLAB>, with all the supplementary material needed to replicate the results. All models have been coded in MATLAB (release R2019a) and solved using the default settings of the ‘linprog’ and ‘fsolve’ optimization functions. They can be run either in the benevolent or aggressive version, and with arithmetic or geometric means as aggregator of the cross-efficiency scores.

To illustrate our comparison of cross-efficiency models from the *empirical perspective*, we used a new database, consisting of 102 warehouses operating in the Benelux area during the period 2012–2017. The data, along with accompanying information on variable codification and survey methods (questionnaire), are available for downloading at <https://doi.org/10.25397/eur.8279426>. It is well known that warehouses are undergoing profound changes from technological, operational, and organizational perspectives. Their success in the market depends on whether they are capable of sustaining high levels of absolute and relative productivity against competitors. Although the warehousing and storage industry in the EU accounted for 73 billion EUR in 2015 and the sector is growing faster than the EUs GDP ([29]), there appears to be a remarkable research gap in analyzing overall efficiency. Thus, beyond providing an illustration to the theory, our numerical exercise is interesting as such.

Finally, from a *managerial perspective*, we compare and evaluate the relative merits of the various methods on a number of dimensions, to provide managerial insights and guidance when choosing among them. The main dimensions are: a) ability to discriminate among warehouses; b) ability to provide consistent and credible rankings for managerial decision making; that is, proximity across methods in terms of statistical differences across the alternative rankings; c) extendability to real-life in-house business applications, implementation ease, and computational requirements; and d) sensitivity of the rankings to scale changes, erroneous entries, and removal of efficient peers from the reference frontier.

The paper unfolds as follows. Section 2 introduces the general idea of cross-efficiency measurement. Section 3 considers the case of qualitative variables. Section 4 surveys the various kinds of secondary goals. Section 5 is concerned with the empirical implementation and comparison of the outcomes of the various cross-efficiency methods. Section 6 contains the managerial insights. Section 7 summarizes and concludes. An on-line Appendix provides auxiliary materials.

2. The basic idea of cross-efficiency measurement

Given input-output data for a set of decision-making units (DMUs, also called firms or production units), a linear DEA program generates for each DMU an efficiency score plus unit-specific weights or shadow prices for all the inputs and outputs. These weights can be used to fill a square matrix of so-called cross-efficiency values, where each unit is appraised by each unit. Averaging those values row- or column-wise delivers aggregate measures for comparing the efficiencies of the production units. However, as is well known, the weights may not be unique, and therefore much of the discussion in the literature is about how appropriate weights can be selected.

Let there be N inputs, the (positive) quantities of which are measured by a vector $x = (x_1, \dots, x_N)$, and M outputs, the (non-negative) quantities of which are measured by a vector $y = (y_1, \dots, y_M)$. Given K observed production units, we have a set of data $\{(x^k, y^k), k = 1, \dots, K\}$.

Using DEA, for each firm $k = 1, \dots, K$ its radial input technical efficiency (ITE), assuming constant returns to scale (CRS), is conventionally calculated by the so-called CCR model²

$$ITE_{CCR}(x^k, y^k) = \min_{\delta, \lambda} \left\{ \delta \mid \sum_{k'=1}^K \lambda_{k'} x^{k'} \leq \delta x^k, y^k \leq \sum_{k'=1}^K \lambda_{k'} y^{k'}, \lambda_{k'} \geq 0, k' = 1, \dots, K \right\}. \quad (1)$$

² Named after [14]. The restriction to the input orientation and CRS is for expository convenience. Under variable returns to scale (VRS), input-orientated cross-efficiency scores may become negative; see [40] for a discussion of this problem. Recently, however, [8,9] have overcome this limitation by proposing a cross-efficiency method in which the input and output multipliers (u^k, v^k) of model (2) below are interpreted as shadow prices of an economic cross-efficiency model à la [31]. This proposal prevents the occurrence of negative cross-efficiency scores under VRS, while offering a ranking of DMUs grounded in economic theory.

As is well known, $ITE_{ccr}(x^k, y^k)$, with values between 0 and 1, is an inverse measure of the distance of firm k to the frontier (= the envelopment of the dataset). Such technical efficiencies are therefore used to compare firms. Put otherwise, firm k is said to be more efficient than firm ℓ if $ITE_{ccr}(x^k, y^k) \geq ITE_{ccr}(x^\ell, y^\ell)$.

Is the use of $ITE_{ccr}(\cdot)$ for comparing efficiencies warranted? To judge this we look at the dual LP problem³

$$ITE_{ccr}(x^k, y^k) = \max_{u, v} \{u \cdot y^k \mid u \cdot y^{k'} - v \cdot x^{k'} \leq 0, v \cdot x^k = 1, u \geq 0, v \geq \epsilon, k' = 1, \dots, K\} \\ = \max_{u, v} \left\{ \frac{u \cdot y^k}{v \cdot x^k} \mid \frac{u \cdot y^{k'}}{v \cdot x^{k'}} \leq 1, u \geq 0, v \geq \epsilon, k' = 1, \dots, K \right\} = \frac{u^k \cdot y^k}{v^k \cdot x^k}, \quad (2)$$

where ϵ represents an infinitesimal lower bound for the multipliers, and (u^k, v^k) is a solution of the maximization problem. It appears that⁴

$$ITE(x^k, y^k) \geq ITE(x^\ell, y^\ell) \Leftrightarrow \frac{u^k \cdot y^k}{v^k \cdot x^k} \geq \frac{u^\ell \cdot y^\ell}{v^\ell \cdot x^\ell}, \quad (3)$$

that is, the comparison of the two firms involves not only their input and output quantities, as one would expect, but also two different vectors of weights, namely (u^k, v^k) and (u^ℓ, v^ℓ) . These shadow prices are, in general, not unique. However, for the time being let us abstract from the nonuniqueness. It would make more sense to base the efficiency comparison of the two firms on the comparison of either

$$\frac{u^k \cdot y^k}{v^k \cdot x^k} \text{ and } \frac{u^k \cdot y^\ell}{v^k \cdot x^\ell}, \text{ or } \frac{u^\ell \cdot y^k}{v^\ell \cdot x^k} \text{ and } \frac{u^\ell \cdot y^\ell}{v^\ell \cdot x^\ell},$$

depending on whether the weights of k or ℓ are used. This constitutes the idea behind the concept of cross-efficiency measurement.

Thus, the cross input technical efficiency (CITE) (score) of firm ℓ with respect to firm k is defined as⁵

$$CITE(x^\ell, y^\ell | k) \equiv \frac{u^k \cdot y^\ell}{v^k \cdot x^\ell} \quad (\ell, k = 1, \dots, K), \quad (4)$$

where (u^k, v^k) satisfies Eq. (2). Notice that $CITE(x^\ell, y^\ell | \ell) = ITE(x^\ell, y^\ell)$ ($\ell = 1, \dots, K$). This could be called the self-appraisal score of firm ℓ . The (arithmetic) mean appraisal score of firm ℓ by all its colleagues is given by $\sum_{k=1, k \neq \ell}^K CITE(x^\ell, y^\ell | k) / (K - 1)$ ($\ell = 1, \dots, K$). The (arithmetic) mean overall appraisal score of firm ℓ , called the cross input technical efficiency (CITE) (score) of firm ℓ , is then given by

$$\sum_{k=1}^K CITE(x^\ell, y^\ell | k) / K \quad (\ell = 1, \dots, K), \quad (5)$$

which is a weighted mean of self-appraisal and colleague-appraisal scores, with weights $1/K$ and $(K - 1)/K$, respectively.⁶ Firm ℓ is now said to be more efficient than firm ℓ' if $\sum_{k=1}^K CITE(x^\ell, y^\ell | k) / K \geq \sum_{k=1}^K CITE(x^{\ell'}, y^{\ell'} | k) / K$, or

$$\frac{\sum_{k=1}^K CITE(x^\ell, y^\ell | k)}{\sum_{k=1}^K CITE(x^{\ell'}, y^{\ell'} | k)} \geq 1.$$

[38] suggest to adjust the CITE scores such that the means of the appraisals by each DMU are the same; that is

$$\sum_{\ell=1}^K CITE(x^\ell, y^\ell | k) / K = \sum_{\ell=1}^K CITE(x^\ell, y^\ell | k') / K \quad (k, k' = 1, \dots, K).$$

Such an adjustment means that the unweighted mean in expression (5) is replaced by a weighted mean. Li et al. [38] interpret those weights as representing some degree of “generosity” from the side of the appraising DMUs. However, this interpretation would only be warranted if “appraisal” is a process in which a DMU is actively involved. As long as there is no explicit relation between the characteristics of a DMU and the shadow prices generated by the LP problem (2), there is no reason to insert some weighting in the definition of the overall appraisal score of a firm.

The interpretation of the measure defined by expression (5) in the single-input (or single-output) case was pointed out by Anderson et al. [5]. When $N = 1$ the vectors x and v become scalars and their inner product reduces to simple multiplication. Then it appears that

$$\frac{1}{K} \sum_{k=1}^K CITE(x^\ell, y^\ell | k) = \frac{1}{x^\ell} \left(\frac{1}{K} \sum_{k=1}^K \frac{u^k}{v^k} \right) \cdot y^\ell \quad (\ell = 1, \dots, K); \quad (6)$$

³ Notation: u and v are vectors of dimension M and N , respectively, and the dot denotes the inner product. From the immediate context it will be clear whether the symbols 0 and 1 designate scalars or vectors of 0s and 1s.

⁴ From here the subscript “ccr” will be deleted as $ITE_{ccr}(\cdot)$ may be replaced by $ITE_{rank}(\cdot)$ as defined below by expression (11).

⁵ The notation is chosen so that the functional structure becomes explicit.

⁶ [42] show remarkable differences between self-appraisal and mean colleague-appraisal scores.

that is, the outputs of each firm are weighted by the same vector of mean (relative) shadow prices, and then the aggregate output quantity is divided by the scalar input quantity. This is a measure of productivity.

In the general case, however, the interpretation is not so straightforward. As mentioned, the inverse of $CITE(x^\ell, y^\ell | \ell)$ measures the distance of firm ℓ to the technological frontier as given by the envelopment of the data. Such a nice interpretation is lacking for $CITE(x^\ell, y^\ell | k)$ for $k \neq \ell$, see [32,45]. Thus, what precisely are we averaging in expression (5), and is the arithmetic mean the only option?

All the appraisers are using different, incommensurable, measuring rods. In expression (5) the arithmetic mean serves as merging function. One could ask whether this is the “best” one. Let $F(a_1, \dots, a_K)$ be a (positive, real-valued) merging function; that is, a function that combines all the appraisal scores into a summary score. The following properties seem required:

- Agreement: $F(a, \dots, a) = a$.
- Symmetry: $F(a_1, \dots, a_K) = F(a_{\pi(1)}, \dots, a_{\pi(K)})$ for any permutation π of $\{1, \dots, K\}$.

We further notice that the scores of each appraiser $k = 1, \dots, K$ constitute a (different) ratio scale, because each ratio $CITE(x^\ell, y^\ell | k) / CITE(x^{\ell'}, y^{\ell'} | k)$ for $\ell, \ell' = 1, \dots, K$ admits a meaningful interpretation, namely as a productivity index (= ratio of output quantity index over input quantity index) of firm ℓ relative to ℓ' . Then [1](Corollary 3.1) show that each ratio of merged scores

$$\frac{F(CITE(x^\ell, y^\ell | 1), \dots, CITE(x^\ell, y^\ell | K))}{F(CITE(x^{\ell'}, y^{\ell'} | 1), \dots, CITE(x^{\ell'}, y^{\ell'} | K))} \quad (\ell, \ell' = 1, \dots, K)$$

is meaningful if and only if $F(\cdot)$ is the geometric mean. Thus, instead of expression (5) one should use

$$\prod_{k=1}^K (CITE(x^\ell, y^\ell | k))^{1/K}. \quad (7)$$

Then firm ℓ is more efficient than firm ℓ' if $\prod_{k=1}^K (CITE(x^\ell, y^\ell | k))^{1/K} \geq \prod_{k=1}^K (CITE(x^{\ell'}, y^{\ell'} | k))^{1/K}$, or

$$\prod_{k=1}^K \left(\frac{CITE(x^\ell, y^\ell | k)}{CITE(x^{\ell'}, y^{\ell'} | k)} \right)^{1/K} \geq 1.$$

At the left-hand side of this inequality we see an unweighted geometric mean of (Lowe-type) productivity indices.⁷

It is interesting to compare this to what happens when the arithmetic mean (5) is used. The ratio of arithmetic mean overall scores can be expressed in two ways, as

$$\begin{aligned} \frac{\sum_{k=1}^K CITE(x^\ell, y^\ell | k)}{\sum_{k=1}^K CITE(x^{\ell'}, y^{\ell'} | k)} &= \sum_{k=1}^K \left(\frac{CITE(x^\ell, y^\ell | k)}{\sum_{k=1}^K CITE(x^\ell, y^\ell | k)} \frac{CITE(x^{\ell'}, y^{\ell'} | k)}{\sum_{k=1}^K CITE(x^{\ell'}, y^{\ell'} | k)} \right) \\ &= \left(\sum_{k=1}^K \frac{CITE(x^\ell, y^\ell | k)}{\sum_{k=1}^K CITE(x^\ell, y^\ell | k)} \left(\frac{CITE(x^\ell, y^\ell | k)}{CITE(x^{\ell'}, y^{\ell'} | k)} \right)^{-1} \right)^{-1}, \end{aligned}$$

thus, as a weighted arithmetic or harmonic mean of (Lowe-type) productivity indices. Now we know that a (weighted) arithmetic mean is greater than or equal to a (weighted) geometric mean, and a (weighted) harmonic mean is less than or equal to a (weighted) geometric mean, but the relation between weighted and unweighted means is uncertain, as being dependent on the covariance between relative efficiencies and productivity changes.

Summarizing, we advocate the geometric mean, expression (7), as a meaningful aggregator of cross-efficiency scores given the properties it satisfies and the fact that, by adopting this functional form, the ratio of cross-efficiency scores of two DMUs can be interpreted as a measure of relative productivity.⁸ Hence, in the empirical application we report geometric means. The MATLAB code accompanying this study, however, calculates also arithmetic means.

3. Dealing with ordinal variables

In the foregoing section it has tacitly been assumed that all the variables are cardinal. In practice this is not always the case. Qualitative inputs or outputs result in variables measured according to some rank order or Likert scale; e.g., ‘priority’ or ‘customer satisfaction’ could be ‘high’, ‘medium’, or ‘low’. While such variables are easily understood by management, they cannot immediately be ascribed a cardinal meaning in the standard DEA framework. We are following here the approach by Cook and Zhu [20].⁹

⁷ See [10] for the definition of Lowe quantity indices. A productivity index is defined as an output quantity index divided by an input quantity index.

⁸ In terms of [53] expression (7) is the geometric mean of optimistic cross-efficiency scores. The focus of this paper is on efficiency rather than cross-efficiency scores. The paper distinguishes between optimistic and pessimistic efficiency scores, and proposes the geometric mean of the two as a ranking device.

⁹ This selection is based on [19](Section 5.4). Performing DEA on non-cardinally measured variables is an area of ongoing research, for which the reader is referred to [60](Chapter 18) and the literature review in the introductory section of [28].

Without loss of generality, let the vectors of input and output quantities be partitioned in the following way: $x = (x_c, x_o) \equiv (x_{c1}, \dots, x_{cP}, x_{o1}, \dots, x_{oQ})$, and $y = (y_c, y_o) \equiv (y_{c1}, \dots, y_{cR}, y_{o1}, \dots, y_{oS})$, with $N = P + Q$ and $M = R + S$, where the subscripts c and o denote cardinal and ordinal, respectively. Then the LP program in expression (2) becomes

$$ITE_{CCR}(x^k, y^k) = \max_{u_c, v_c, u_o, v_o} \left\{ \frac{u_c \cdot y_c^k + u_o \cdot y_o^k}{v_c \cdot x_c^k + v_o \cdot x_o^k} \mid \frac{u_c \cdot y_c^{k'} + u_o \cdot y_o^{k'}}{v_c \cdot x_c^{k'} + v_o \cdot x_o^{k'}} \leq 1, u_c, u_o \geq 0, v_c, v_o \geq \epsilon, k' = 1, \dots, K \right\} = \frac{u_c^k \cdot y_c^k + u_o^k \cdot y_o^k}{v_c^k \cdot x_c^k + v_o^k \cdot x_o^k}, \quad (8)$$

where now $(u_c^k, v_c^k, u_o^k, v_o^k)$ is a solution to the maximization problem.

The first step is to code each qualitative variable numerically in such a way that the various rank positions are assigned scores following the rule ‘the higher the better’ for outputs and ‘the higher the worst’ for inputs; e.g., the score 3 is assigned to ‘high’, the score 2 to ‘medium’, and the score 1 to ‘low’. It is tempting to use these scores as quantities and then solve the LP program (8). We call this ‘simple cardinalization’.

Though this is routinely practised, it basically constitutes a neglect of the ordinal character of the scores. The point is that the differences between two scores of the same variable are not in any way commensurable; e.g., $4 - 3$ is not ‘the same’ as $2 - 1$. And the same score on two different variables is not commensurable either.

The procedure to cardinalize ordinal variables is to split each qualitative variable in as many subvariables as there are rank positions and to assign a certain, as yet undetermined, ‘worth’ to each of those subvariables. Formally, the ‘worth’ for rank position ℓ of output variable oj is denoted by $y_{oj\ell}$ ($j = 1, \dots, S; \ell = 1, \dots, L_j$), and the ‘worth’ for rank position ℓ of input variable oi is denoted by $x_{oi\ell}$ ($i = 1, \dots, Q; \ell = 1, \dots, L_i$). Finally, dummy variables are created such that $\gamma_{oj\ell}^k = 1$ if DMU k has qualitative output attribute j with rank greater than or equal to ℓ , and $= 0$ otherwise ($k = 1, \dots, K$), ($j = 1, \dots, S; \ell = 1, \dots, L_j$), and $\gamma_{oi\ell}^k = 1$ if DMU k has qualitative input attribute i with rank greater than or equal to ℓ , and $= 0$ otherwise ($k = 1, \dots, K$), ($i = 1, \dots, Q; \ell = 1, \dots, L_i$).

Output and input quantities are then defined by

$$y_{oj}^k \equiv \sum_{\ell=1}^{L_j} \gamma_{oj\ell}^k y_{oj\ell} \quad (k = 1, \dots, K, j = 1, \dots, S) \quad (9)$$

$$x_{oi}^k \equiv \sum_{\ell=1}^{L_i} \gamma_{oi\ell}^k x_{oi\ell} \quad (k = 1, \dots, K, i = 1, \dots, Q). \quad (10)$$

Notice that by setting all the ‘worth’s’ equal to 1 we are back to the simple cardinalization mentioned above. Substituting the quantities defined by expressions (9) and (10) into expression (8), and defining new weights by

$$\theta_{oj\ell} \equiv u_{oj} y_{oj\ell} \quad (j = 1, \dots, S; \ell = 1, \dots, L_j)$$

$$\phi_{oi\ell} \equiv v_{oi} x_{oi\ell} \quad (i = 1, \dots, Q; \ell = 1, \dots, L_i),$$

we obtain

$$\begin{aligned} ITE_{rank}(x^k, y^k) &= \max_{u_c, v_c, \theta_o, \phi_o} \left\{ \frac{\sum_{j=1}^S u_{cj} y_{cj}^k + \sum_{j=1}^S \sum_{\ell=1}^{L_j} \theta_{oj\ell} \gamma_{oj\ell}^k}{\sum_{i=1}^Q v_{ci} x_{ci}^k + \sum_{i=1}^Q \sum_{\ell=1}^{L_i} \phi_{oi\ell} \gamma_{oi\ell}^k} \mid \frac{\sum_{j=1}^S u_{cj} y_{cj}^{k'} + \sum_{j=1}^S \sum_{\ell=1}^{L_j} \theta_{oj\ell} \gamma_{oj\ell}^{k'}}{\sum_{i=1}^Q v_{ci} x_{ci}^{k'} + \sum_{i=1}^Q \sum_{\ell=1}^{L_i} \phi_{oi\ell} \gamma_{oi\ell}^{k'}} \leq 1, u_c, v_c \geq \epsilon_c, \theta_o, \phi_o \geq \epsilon_o, k' = 1, \dots, K \right\} \\ &= \frac{\sum_{j=1}^S u_{cj}^k y_{cj}^k + \sum_{j=1}^S \sum_{\ell=1}^{L_j} \theta_{oj\ell}^k \gamma_{oj\ell}^k}{\sum_{i=1}^Q v_{ci}^k x_{ci}^k + \sum_{i=1}^Q \sum_{\ell=1}^{L_i} \phi_{oi\ell}^k \gamma_{oi\ell}^k}. \end{aligned} \quad (11)$$

This program has the same structure as the program in expression (8), but with a much larger number of variables and hence dimensions. Thus the computational burden is higher, which is why, despite its shortcomings, simple cardinalization is usually preferred to the more complex rank-order model outlined above. To assess the empirical impact of the cardinalization choice, in Section 5 we juxtapose the scores obtained by simple cardinalization implementations and those based on a rank-order model. Notice that in expression (11) different lower bounds have been introduced for the cardinal and ordinal variables, ϵ_c and ϵ_o , respectively. This is to accommodate the importance of the various ordinal variables – in the empirical section about warehouse performance we comment on the values that have been adopted.

4. The nonuniqueness problem

As noted below expression (2), the weight vectors (u^k, v^k) ($k = 1, \dots, K$) are not unique. For instance, all the extreme efficient units have an infinite number of optimal weights, as do all the inefficient units belonging to or projected onto the weak efficiency frontier, for which the optimal solution involves positive slacks. Retracing the steps taken in the previous section, this means that neither the cross-efficiencies $CITE(x^\ell, y^\ell | k)$ ($\ell, k = 1, \dots, K$) nor their means $\sum_{k=1}^K CITE(x^\ell, y^\ell | k) / K$ or $\prod_{k=1}^K (CITE(x^\ell, y^\ell | k))^{1/K}$ ($\ell = 1, \dots, K$) are unique.

The literature provides us with a number of approaches to obtain (approximately) unique scores. We discuss them under three headings, roughly corresponding to their genesis in time.

4.1. Weight selection approaches: aggressive and benevolent

The idea behind the first approach is to select those vectors from the set of all the optimal weights solving the LP problem (2) which have the greatest discriminatory power. Based on [48], [26,27] considered a number of options. The most natural is based on the $CITE(x^\ell, y^\ell | k)$ ratio formulation in expression (4):

$$(u^{k*}, v^{k*}) \equiv \arg \min_{u^k, v^k} \left\{ \frac{1}{K-1} \sum_{k'=1, k' \neq k}^K CITE(x^{k'}, y^{k'} | k) \mid ITE(x^k, y^k) = \frac{u^k \cdot y^k}{v^k \cdot x^k} \right\}. \quad (12)$$

For each $k = 1, \dots, K$ the pair (u^{k*}, v^{k*}) is then used to compute cross input technical efficiencies according to expression (4).

The secondary goal set by the minimization problem in expression (12) is to obtain the greatest difference between the self-appraisal score of firm k and the mean of the appraisals by k of all its rivals. This is a highly nonlinear, fractional problem, which at the end of the previous century was still deemed unsolvable.

As a feasible alternative [48] had already considered

$$(u^{k*}, v^{k*}) \equiv \arg \min_{u^k, v^k} \left\{ u^k \cdot \bar{y}^k - v^k \cdot \bar{x}^k \mid ITE(x^k, y^k) = \frac{u^k \cdot y^k}{v^k \cdot x^k} \right\}, \quad (13)$$

where $(\bar{x}^k, \bar{y}^k) \equiv \sum_{k'=1, k' \neq k}^K (x^{k'}, y^{k'})$. Thus, from all the pairs of shadow price vectors generated by the LP problem (2) the pair is selected which minimizes the profit of the aggregate, k -excluded, production unit. Given the linear objective function, this approach has become the preferred method in cross-efficiency applications.

The outcome, however, does depend on the size distribution of the firms. This can be seen by noticing that in expression (13) above $u^k \cdot \bar{y}^k - v^k \cdot \bar{x}^k = v^k \cdot \bar{x}^k \times (CITE(\bar{x}^k, \bar{y}^k | k) - 1)$. To overcome this problem, Doyle and Green [26,27] considered

$$(u^{k*}, v^{k*}) \equiv \arg \min_{u^k, v^k} \left\{ CITE(\bar{x}^k, \bar{y}^k | k) \mid ITE(x^k, y^k) = \frac{u^k \cdot y^k}{v^k \cdot x^k} \right\}. \quad (14)$$

Here, from all the pairs of weights generated by the LP problem (2) the pair is selected which minimizes the cross-efficiency of the aggregate, k -excluded, production unit. Put otherwise, the mean of ratios in problem (12) is replaced by the ratio of means in problem (14).

Because of the minimization operator in the above three expressions the methods were classified as 'aggressive'. Replacing the min by the max operator turns the 'aggressive' methods into 'benevolent' ones.

Finally, as mentioned in the Introduction, there are more recent proposals within the weight selection approach. Among these, in chronological order, those proposed by Contreras [18], Jahanshahloo et al. [35], Maddahi et al. [44], Wu et al. [54], [58], Zohrehbandian and Gavgani [61]. All these models represent alternative secondary goals that place different restrictions on the weights. For the purpose of our paper, we let the entire class of models be represented by the initial proposals of [48]: the *classic* model of expression (12), and the *linear* model of expression (13).

4.2. The multiplicative approach

[14] introduced the multiplicative variant of the LP problem (2):

$$ITE'(x^k, y^k) \equiv \max_{u, v} \left\{ \frac{\prod_{m=1}^M (y_m^k)^{u_m}}{\prod_{n=1}^N (x_n^k)^{v_n}} \mid \frac{\prod_{m=1}^M (y_m^{k'})^{u_m}}{\prod_{n=1}^N (x_n^{k'})^{v_n}} \leq 1, u \geq 1, v \geq 1, k' = 1, \dots, K \right\} = \frac{\prod_{m=1}^M (y_m^k)^{u_m^k}}{\prod_{n=1}^N (x_n^k)^{v_n^k}} \quad (15)$$

where (u^k, v^k) is a solution of the maximization problem. By taking logarithms, this appears to be a CRS additive DEA model. The counterpart of expression (4) then becomes

$$CITE'(x^\ell, y^\ell | k) \equiv \frac{\prod_{m=1}^M (y_m^\ell)^{u_m^k}}{\prod_{n=1}^N (x_n^\ell)^{v_n^k}} \quad (\ell, k = 1, \dots, K), \quad (16)$$

where (u^k, v^k) satisfies Eq. (15). Cook and Zhu [21], [22] proposed to merge these scores by a geometric mean, and defined

$$\max_{u^k, v^k} \left\{ \prod_{k=1}^K (CITE'(x^\ell, y^\ell | k))^{1/K} \mid ITE'(x^k, y^k) = \frac{\prod_{m=1}^M (y_m^k)^{u_m^k}}{\prod_{n=1}^N (x_n^k)^{v_n^k}}, k = 1, \dots, K \right\} \quad (17)$$

as the final efficiency score of DMU $\ell = 1, \dots, K$. By taking logarithms this maximization problem becomes linear, though its number of constraints may be considerable (namely K^2).

The fact that the function defined by expression (15) is not invariant to changes in the units of measurement of the inputs and outputs might be seen as a problem. Charnes et al. [14] provided a solution by considering a slight modification of the foregoing maximization problem, namely by inserting a scalar ω so that

$$ITE''(x^k, y^k) \equiv \max_{u, v, \omega} \left\{ e^{\omega} \frac{\prod_{m=1}^M (y_m^k)^{u_m}}{\prod_{n=1}^N (x_n^k)^{v_n}} \mid e^{\omega} \frac{\prod_{m=1}^M (y_m^{k'})^{u_m}}{\prod_{n=1}^N (x_n^{k'})^{v_n}} \leq 1, u \geq 1, v \geq 1, k' = 1, \dots, K \right\} = e^{\omega^k} \frac{\prod_{m=1}^M (y_m^k)^{u_m^k}}{\prod_{n=1}^N (x_n^k)^{v_n^k}}, \quad (18)$$

where (u^k, v^k, ω^k) is a solution of the maximization problem. By taking logarithms, this appears to be a VRS additive DEA model [50]. pointed out an additional benefit of this model: there is “no longer any fear of zero (or infinitesimal epsilon) weights.”

4.3. The game-theoretic approach

The three Doyle and Green variants as well as the multiplicative Cook and Zhu method are essentially two-step algorithms, as appears from the definitional equations. Liang et al. [39] took another route. These authors considered the following modification of the LP problem (2):

$$\max_{u,v} \left\{ \frac{u \cdot y^k}{v \cdot x^k} \mid \frac{u \cdot y^{k'}}{v \cdot x^{k'}} \leq 1, \frac{u \cdot y^d}{v \cdot x^d} \geq \alpha^{dt}, u \geq 0, v \geq 0, k' = 1, \dots, K \right\} = \frac{u^k(\alpha^{dt}) \cdot y^k}{v^k(\alpha^{dt}) \cdot x^k}, \quad (19)$$

where the shadow price vectors $(u^k(\alpha^{dt}), v^k(\alpha^{dt}))$ solve the maximization problem, $d = 1, \dots, K$, and t is some auxiliary label. The additional constraint means that the cross input technical efficiency of firm d with respect to firm k should be above some level α^{dt} . Obviously, all the weights are then functions of this level. Notice that the constraints imply that $\alpha^{dt} \leq 1$ ($d = 1, \dots, K$).

If the vector pair $(u^k(\alpha^{dt}), v^k(\alpha^{dt}))$ solves the maximization problem (19) then

$$\frac{u^k(\alpha^{dt}) \cdot y^d}{v^k(\alpha^{dt}) \cdot x^d} \geq \alpha^{dt} \text{ and } \frac{u^k(\alpha^{dt}) \cdot y^{k'}}{v^k(\alpha^{dt}) \cdot x^{k'}} \leq 1 \quad (k' = 1, \dots, K).$$

The second inequality, however, implies that $(u^k(\alpha^{dt}), v^k(\alpha^{dt}))$ satisfies the conditions defining the maximization problem (2) for $ITE(x^d, y^d)$, and thus

$$\frac{u^k(\alpha^{dt}) \cdot y^d}{v^k(\alpha^{dt}) \cdot x^d} \leq ITE(x^d, y^d).$$

Combining this with the first of the previous two inequalities leads to the conclusion that feasibility of the maximization problem (19) implies that $\alpha^{dt} \leq ITE(x^d, y^d)$.

Thus, let $\alpha^{dt} \leq ITE(x^d, y^d)$ for $d = 1, \dots, K$. Expression (2) then tells us that there exist (u^d, v^d) such that

$$\alpha^{dt} \leq \frac{u^d \cdot y^d}{v^d \cdot x^d} \text{ and } \frac{u^d \cdot y^{k'}}{v^d \cdot x^{k'}} \leq 1 \quad (k' = 1, \dots, K).$$

This, however, means that (u^d, v^d) satisfies the conditions defining the maximization problem (19), and hence

$$\frac{u^k(\alpha^{dt}) \cdot y^k}{v^k(\alpha^{dt}) \cdot x^k} \geq \frac{u^d \cdot y^k}{v^d \cdot x^k} = CITE(x^k, y^k | d) \quad (k = 1, \dots, K),$$

where the last step rests on definition (4). Taking the arithmetic mean over both sides of this inequality leads to

$$\frac{1}{K} \sum_{d=1}^K \frac{u^k(\alpha^{dt}) \cdot y^k}{v^k(\alpha^{dt}) \cdot x^k} \geq \frac{1}{K} \sum_{d=1}^K CITE(x^k, y^k | d) \quad (k = 1, \dots, K).$$

Now, given a set of levels $\{\alpha^{dt}; d = 1, \dots, K\}$ it is rather natural to define for each $k = 1, \dots, K$ the next level as

$$\alpha^{k,t+1} \equiv \frac{1}{K} \sum_{d=1}^K \frac{u^k(\alpha^{dt}) \cdot y^k}{v^k(\alpha^{dt}) \cdot x^k} \quad (k = 1, \dots, K); \quad (20)$$

that is, as the mean input technical efficiency of DMU k such that the cross efficiency of each DMU d does not drop below the level α^{dt} .

Basically, expression (20) defines a mapping from the set $[0, 1]^K$ to $[0, 1]^K$. For $\alpha^{dt} \in [\sum_{k=1}^K CITE(x^d, y^d | k)/K, ITE(x^d, y^d)]$ ($d = 1, \dots, K$) [39] showed that this mapping is continuous. Thus, by Brouwer's Fixed Point Theorem, there exists a vector of α 's such that

$$\alpha^k = \frac{1}{K} \sum_{d=1}^K \frac{u^k(\alpha^d) \cdot y^k}{v^k(\alpha^d) \cdot x^k} \quad (k = 1, \dots, K).$$

Liang et al. [39] also showed that the iterative system built on expression (20) converges. Initially, the levels are thereby chosen as

$$\alpha^{\ell 0} = \sum_{k=1}^K CITE(x^\ell, y^\ell | k)/K \quad (\ell = 1, \dots, K); \quad (21)$$

that is, the mean cross input technical efficiencies generated by the original DEA problems (2). Then the levels $\alpha^{\ell 1}$ ($\ell = 1, \dots, K$) are generated by applying expression (20), *etcetera*, until convergence is reached.

Table 1
Descriptive statistics of input and output variables, 2017.

	Input				Output			
	FTEs	Floor space (m ²)	SKUs	Automation score	Order lines	Special processes	Error free %	Order flexibility
Minimum	5	500	100	2	54	2	1	12
Median	30	9250	4600	6	1200	6	7	22
Average	59	18,244	21,088	7	4931	6	6	21
Maximum	350	275,000	400,000	16	55,000	10	9	30
Standard deviation	74	32,414	57,393	3	9815	2	2	4
Growth rate 2012-17 (%)	14.2	37.2	11.8	38.1	34.2	24.5	39.6	18.2

Though the final levels cannot be considered as DEA-based mean cross-efficiencies, they can be considered as Nash equilibrium outcomes of a non-cooperative game in which the firms are the players. Hence the name ‘Game Cross Efficiency’, as coined by Liang et al. [39]. As already remarked in the Introduction, Wu et al. [56] developed an alternative model from the perspective of cooperative games, while Wu et al. [57] went further by using the perspective of a bargaining game. Later on, Ramón et al. [47], Wang and Chin [52] let the weights for each unit be determined without considering the impact on its rivals. Finally, the game-theoretic approach has been further extended and applied by Ma et al. [43].

5. Comparing the methods on warehouse data

5.1. Survey methods and variable selection

For DEA models to have sufficient discriminatory power and provide meaningful rankings, high degrees of freedom are necessary, with the balance between observations and variables playing a pivotal role. To illustrate our comparison of cross-efficiency methods we used a database consisting of 102 warehouses, whose operations are characterized in terms of four inputs and four outputs. As [24,25,30] extensively surveyed and analyzed warehouses, we rely on their choice of input and output variables to characterize the warehouse processes.

The database is the result of a comprehensive online outreach and response collection process executed in 2012 and 2017. A detailed discussion of the survey methods, the data collection, and the input and output variables can be found in [11].¹⁰ Input variables are: 1) *Number of full time equivalent employees* (FTEs); 2) *Warehouse size in m²* (Floor space); 3) *Number of stock keeping units* (SKUs); and 4) *Level of automation* (Automation score). The last variable captures the stock of automation technology implemented in the warehouse and is defined as the sum of several hardware and software automation technologies. Output variables are: 1) *Number of daily order lines* (Order lines); 2) *Number of special processes*; 3) *Error-free order line percentage* (Error free %), measured on an increasing nine-point ordinal scale; and 4) *Order flexibility*, measured on a thirty-point ordinal scale.

The last two output variables are of qualitative nature. Hence, there is sufficient reason to not only apply the conventional CCR model, expression (2), after the qualitative variables have undergone simple cardinalization, but also to apply the full rank-order model, expression (11), on all the variables as they are. A comparison of their outcomes allows us to study whether these models yield results that are statistically different.

Table 1 presents the descriptive statistics for the input and the output variables in 2017. The last row shows the growth rate from 2012 to 2017, portraying a significant increase in the scale of operations as the average number of SKUs increased by more than 10%, and floor space and automation by almost 40%. These trends show that growth goes hand-in-hand with both hardware and software investments, and with the substitution of labor by capital (e.g., robotics). The growth is also observed on the output side. To the extent that on average output quantities have increased more than input quantities we may conclude that the industry has shown productivity growth. Notice that order lines as well as the percentage of error free order lines increased by almost 40%.

5.2. Results

We now present the results of applying the cross-efficiency methods discussed in the foregoing section. Using simple cardinalization of the qualitative variables, we implemented the Sexton-classic method as formulated by expression (12), the Sexton-linear method as formulated by expression (13),¹¹ the multiplicative method as formulated in expression (17),

¹⁰ Appendices A and B provide a summary. The data, along with accompanying information on variable codification and survey methods (questionnaire), are available for downloading at <https://doi.org/10.25397/eur.8279426>.

¹¹ We also implemented the Sexton-ratio method as defined by expression (14), and obtained minor differences with the Sexton-linear method. For this reason the ratio-method results are not reported. Also, throughout the text we make reference to the 2017 results, while their 2012 counterparts are recalled whenever necessary for temporal comparisons. All the individual results of the methods in both years are available upon request.

Table 2
Cross-efficiency scores, 2017.

	Classic on simple cardinalization	Linear	Multiplicative	Game	Classic on rank-order model
Minimum	0.063	0.055	0.000	0.155	0.044
Average	0.309	0.309	0.025	0.506	0.311
Maximum	0.856	0.848	0.911	1.000	0.925
Standard deviation	0.177	0.183	0.110	0.219	0.209

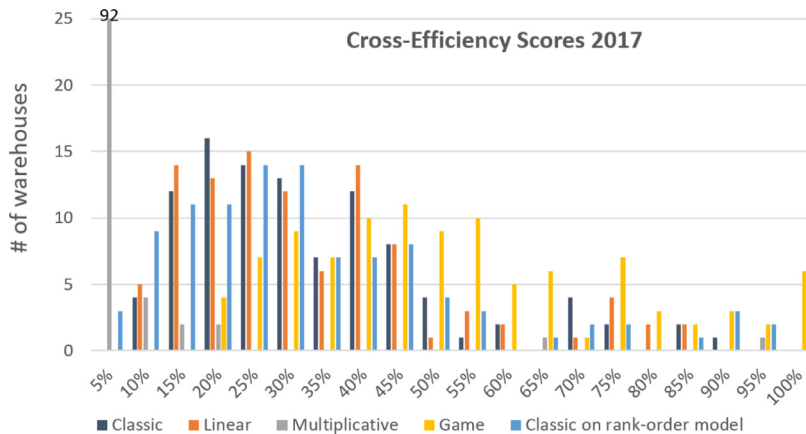


Fig. 1. Cross-efficiency score distribution per method, 2017.

and the game-theoretic method as represented by expression (21). Our fifth set of results corresponds to the Sexton-classic method after the standard model (2) was replaced by the rank-order model (11). All the models were run in their benevolent version, maximizing the peer-appraisal scores, which makes results comparable and in accordance with the market intuition that competitors maximize their own efficiency given a set of weights and constraints. Table 2 reports the summary statistics of the cross-efficiency scores. Following the discussion in Section 2, geometric means are reported. However, we have also calculated the arithmetic means. The correlations between geometric and arithmetic means are between 0.97–1.00 for 2012 and 2017 for all methods except the rank-order based method, where the correlations are 0.69 and 0.72.

While the two Sexton-based methods result in scores of the same order of magnitude per warehouse (as expected given the similarity of the models), the cross-efficiency scores obtained by the multiplicative method are significantly lower (average equal to 0.309 for the first two models and 0.025 for the multiplicative method, with standard deviations 0.177, 0.183, and 0.110, respectively). The game-theoretic method resulted in the highest average score of 0.506 (standard deviation 0.219), and finished after 32 iterations. The scores obtained from the classic method applied to the rank-order model come close to the Sexton-based scores but are consistently higher as a result of the much larger number of variables included in expression (11). Specifically, for *Error-free order line percentage* as many as 8 variables are added, corresponding to the ranking positions between the minimum and maximum values of 1 and 9, respectively. For *Order flexibility* as many as 18 variables are added to the model, corresponding to a range between 12 and 30.

Regarding the rank-order model, we have tried alternative lower bounds for the multipliers corresponding to the set of binary restrictions associated with the ordinal variables, ϵ_0 . Feasible and stable solutions yielding sensible rankings when compared to their cardinal counterparts (and preventing that most of the warehouses are deemed efficient) are obtained for values as low as $1e7$, which is reassuring as it does not impose large values on the multipliers. However, the method is sensitive to the specification of the bounds and may require re-scaling of inputs or outputs. For our unscaled 2017 data, a bound of $1e5$ appears infeasible, and the results of $1e10$ appear not stable with current MATLAB solvers. In addition, the existence of lower bounds may lead to many fully efficient DMUs in models with a larger number of ordinal levels. See [20](1031) for more information on the choice of weight restrictions.

Using the one-sample Kolmogorov-Smirnov (KS) test, the normal distribution assumption could be rejected at $p < 0.01$. The score distributions are visually depicted in Fig. 1. Most scores of the multiplicative method are below 0.01 (92 out of 102 observations). This is probably due to the exponential nature of the calculation, heavily emphasizing efficient warehouses over inefficient ones, which leads to stark score differences compared to additive DEA formulations. We also observe that the game-theoretic method identifies several warehouses as (almost) efficient.

Next we tested whether the five methods yield the same or different distributions from a statistical perspective. Based on the Wilcoxon signed rank test, the hypotheses that the 2017 scores are from the same distribution were rejected for all

Table 3

Top 10 warehouses, by cross-efficiency score and method, 2017.

	Classic on simple cardinalization	Linear	Multiplicative	Game	Classic on rank-order model
1	#098	#098	#067	#049	#098
2	#050	#050	#050	#098	#050
3	#028	#028	#028	#050	#123
4	#104	#066	#027	#028	#058
5	#066	#104	#098	#104	#028
6	#049	#067	#066	#066	#115
7	#067	#049	#104	#027	#049
8	#027	#027	#006	#067	#067
9	#041	#041	#115	#107	#059
10	#107	#107	#040	#041	#041

Table 4Kendall's τ for cross-efficiency rankings, 2017.

	Classic on simple cardinalization	Linear	Multiplicative	Game	Classic on rank-order model	ITE	Super Efficiency
Classic	1						
Linear	0.97	1					
Multiplicative	0.67	0.66	1				
Game	0.76	0.76	0.65	1			
Classic*	0.56	0.56	0.52	0.46	1		
ITE	0.49	0.49	0.52	0.71	0.32	1	
Super Efficiency	0.49	0.49	0.53	0.71	0.32	0.94	1

Note: All p -values are < 0.01 . Classic* = Classic based on rank-order model.

pairs except the two Sexton-based methods. It is then clear that the choice of method is not neutral, leading to different rankings and presentations of warehouse performance within the industry.

Of course, the most critical question is whether the various methods result in comparable rankings. After all, managers are less interested in particular cross-efficiency scores, but rather in how their facilities behave comparatively, and who is best-in-class. Table 3 reports the ranking of the top 10 warehouses by cross-efficiency score for the five methods. Warehouses in bold signify that they are in the top 10 by all five methods.

It is noteworthy that the top 3 are not identical among the five methods and that only four warehouses in 2017 and five in 2012 were ranked in the top 10 by all the methods. Still, the two Sexton-based methods exhibit very similar rankings. The game-theoretic method ranks the top 10 similar to the Sexton-based methods (nine out of ten facilities in the top 10 are the same), whereas the ranking by the multiplicative and the classic-based-on-rank-order-model methods differ already in the first warehouses.

Kendall's τ correlations were calculated for all method pairs. The two Sexton-based methods show correlations of over 0.93 in both years, indicating almost identical rankings. Similarly, the Sexton-based methods and the game-theoretic method correlate by more than 0.73 in both years. The correlations between the multiplicative method and the other four methods are considerably lower, namely in the range 0.52–0.67. The correlation between the cross-efficiency rankings and the first-stage ITE(.) ranking according to expression (2) is also presented. As the latter yields 26 efficient warehouses, an additional ranking was compiled, based on super-efficiency scores, following [4]. The correlations with these two rankings follow a similar pattern. The lowest corresponds to the classic-based-on-rank-order-model method (0.32), followed by the relatively low multiplicative and Sexton-based methods (0.53, 0.49, 0.49), and then by the moderately large game-theoretic method.

6. Managerial insights

This section focuses on the relative merits of the five cross-efficiency measurement methods from a managerial perspective. We hope to facilitate a wider audience's engagement with these methods by summarizing our experience in dealing with the various approaches. With this in mind, we appraised every method with a score from 1 (worst) to 5 (best) on each of the dimensions in Table 5.

This procedure is useful for comparing methods across single dimensions, but is not intended for computing an aggregate score per method, as different situations may lead to different choices. Although fairly standard, we acknowledge that other dimensions can be considered and that our evaluation scores are necessarily subjective. Nevertheless we believe they form a sensible starting point, and we invite practitioners to carry forward this endeavor through their own analyses. Before performing the evaluation we emphasize that the comparison is meaningful since we have executed Wilcoxon signed rank tests to determine whether the sets of results, obtained through the various methods, are significantly different.

Table 5
Comparison dimensions for the cross-efficiency methods.

Dimension	Description
Methodological proximity to standard DEA	Tests how close a cross-efficiency method follows the logic of the familiar first-stage DEA model. As the DEA scores are a natural upper limit for the cross-efficiency scores, methodological proximity makes the results more reliable.
Implementational ease	Tests how easily a method can be implemented and assesses the degree of computational strain when handling a large dataset. Especially for application in non-academic fields, easy implementation and swift calculation in case of automated or frequent application are relevant.
Extendability	Tests how many modifications to the basic model are proposed in the literature, which allows tailoring the method to individual project requirements.
Discriminatory properties	Tests how large the relative differences in cross-efficiency scores are across methods, for those observations that are signaled as efficient at the first DEA stage. As a cross-efficiency method is mainly applied to break ties between efficient observations, this is of great relevance for practitioners.
Sensitivity to changes of scale	Tests how robust the results of a particular method are to scale changes of input and output variables. In volatile environments a low sensitivity to scale changes increases the validity of the results.
Sensitivity to erroneous data	Tests how robust the results of a particular method are to (random) changes in the magnitudes of some inputs and outputs. A low sensitivity to erroneous data increases the reliability of the results, especially when the data is subjective/opinion-dependent, based on estimates, or exposed to human error.
Sensitivity to peers deletion	Tests how robust the results of a particular method are to deleting observations lying on a DEA frontier hyperplane. A low sensitivity to deleting such benchmarks increases the reliability of the results, especially for industrial comparisons, when extreme observations might be part of the sample.

Table 6
Ratings of cross-efficiency methods across dimensions.

	Classic on simple cardinalization	Linear	Multiplicative	Game	Classic on rank-order model
Proximity to plain DEA	4	3	2	5	3
Implementational ease	4	5	4	2	3
Extendability	4	4	2	3	2
Discriminatory properties	5	5	2	4	2
Sensitivity to changes of scale	5	5	2	5	4
Sensitivity to erroneous data	5	5	3	5	5
Sensitivity to peers deletion	4	4	1	3	4

Table 6 contains the results. Explanatory notes can be found in Appendix C. Our conclusion is that the two Sexton-based methods (employing simple cardinalization of the qualitative variables) appeared to perform best from a cost-benefit perspective; i.e., the ability to benchmark firms in a straightforward and reliable manner. This evaluation is based mainly on implementation ease, discriminatory (and therefore ranking) properties, and robustness. Also, despite its computational difficulties, as the rank-order method is the only one accounting for ordinal variables, we recommend its calculation to complement the results obtained with the other methods.

7. Summary and conclusions

In this article we examined cross-efficiency measurement from the perspective of index number theory. Cross-efficiency scores represent measures of relative performance that can be confidently used for bilateral comparisons of productivity. We made the case for the aggregation of elementary cross-efficiency scores by geometric rather than arithmetic means. Based on reviews of the main cross-efficiency methods, we identified four representatives for a study of warehouse performance. Our dataset consists of a sample of 102 warehouses operating in the Benelux area in 2012–2017, whose technology is characterized by four inputs and four outputs. As two of the output variables are essentially qualitative, we extended the rank-order DEA model introduced by Cook and Zhu [20] to be employed in cross-efficiency methods.

We found that the choice of cross-efficiency methods results in statistically different distributions of the efficiency scores as well as different rankings of the warehouses. Indeed there are differences of an order of magnitude between the scores of the two Sexton-based methods (that exhibit the same average value at 0.309) and those of the multiplicative method (where the average is 0.025). The game-theoretic method results in the highest average score of 0.506. Although the rank-order model based on the classic approach stands aside from a methodological perspective, its scores are comparable with those

obtained through the Sexton-based methods, though consistently higher. This result is expected given the larger number of variables included in expression (11).

Focusing on the possibility of managers actually implementing these methods for benchmarking, our findings show that the Sexton-based methods, followed by the game-theoretic method, are superior to the multiplicative method across almost all dimensions of comparison. Choosing between the two Sexton-based methods depends on the preference of the user and the context of the application. Although the ratio and linear methods are slightly more quickly solvable, the classic framework approximates the first-stage DEA program methodologically closest. The game-theoretic method is disadvantageous mainly in terms of required computational time; it also occasionally identifies more than one DMU as fully efficient, which results in ties at the top position. However, growing computer power should be able to compensate for this in the coming years. The rank-order model methodologically differs from the rest by accommodating qualitative variables. Using this model adds robustness to the results obtained with other methods, though the sensitivity to the multiplier bounds is a matter of concern. Overall, given the complexity of implementing the model and its numerical unreliability, we discourage practitioners from resorting to this specific proposal, and explore alternatives such as those considered in [28].

Let us conclude by identifying some empirical limitations of this study and by indicating areas for further research. From the methodological perspective, our conclusions suggest that rather than making further refinements in complex elaborations of secondary goals, academics interested in cross-efficiency methods should shift their attention to some of the fundamental limitations of the methodology. One of the current challenges is the relaxation of the assumption of CRS, which in many situations is an unwarranted technological assumption. The problem is that extending the basic CRS model, as considered in this study, to its VRS counterpart, see [22], may result in negative cross-efficiency scores. Lim and Zhu [40], Wu et al. [55] proposed different work-arounds to this problem. The first authors restrict the value of the numerator of the VRS multiplier formulation of the traditional input-orientated BCC model to be positive, which ensures that cross-efficiencies cannot be negative. The second rely on a geometrical solution and implement a Cartesian coordinate-system translation before solving the model. Aparicio and Zofío (AZ) [8,9], propose an innovative approach to calculate cross-efficiencies, departing from the optimal input weights. These multipliers are interpreted as shadow prices, and it is shown that, under input homotheticity, the bilateral VRS cross-efficiency model is equivalent to the cost-efficiency model of Farrell [31]. In this approach VRS cross-efficiency can be reinterpreted as the cost efficiency of the DMU under evaluation, considering the set of optimal multipliers of the remaining DMUs as reference prices. AZ coin the term Farrell cross-efficiency and extend their results to the notion of Nerlovian cross-inefficiency, where the directional distance function is used as efficiency measure. Besides solving the problem of negative scores under VRS, AZ's approach allows to decompose economic cross-efficiency into technical and allocative efficiency components. Overall, the most salient feature of AZ's proposal, which brings together the cross-efficiency and economic efficiency literatures, is that cross-efficiency is provided with a solid foundation in economic theory. A comparison of methods like those reported in this paper, both under CRS and VRS, would shed light on the role that scale efficiency and allocative efficiency play in the cross-efficiency performance of DMUs.

A further topic is the introduction of the temporal dimension in cross-efficiency measurement. For a first definition of cross-productivity Malmquist indices and Luenberger indicators, to disentangle the role of efficiency change and technological change, see [7]. Future studies could also follow new frameworks which allow more realistic representations of warehouse processes, such as network DEA (see [37], or [46]), in which some variables (for instance, automation) may be treated as an intermediate rather than an input or output.

From an empirical perspective, the degree ('quantity') of automation remains an elusive variable, normally defined by an ordinal scale. A combination of available questions, experience, and expert judgment was used to develop the automation section of our questionnaire. Finally, although a dataset of over 100 warehouses fulfills all minimum requirements for DEA (and exceeds the numbers one sees in other studies of the industry), a larger dataset would be preferable to assess the computational performance of the different cross-efficiency models. Here, a web-hosted solution like the one proposed by Johnson and McGinnis [36] would provide warehouse stakeholders with an interface to submit operational data and have their warehouse ranked through a cross-efficiency computer code to incentivize submission.

Acknowledgements

José L. Zofío thanks the financial support from the Spanish Ministry of Science and Innovation (Ministerio de Ciencia e Innovación), the State Research Agency (Agencia Estatal de Investigación) and the European Regional Development Fund (Fondo Europeo de Desarrollo Regional) under grants EIN2020-112260 and PID2019-105952GB-I00 (AEI/FEDER, UE).

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:[10.1016/j.amc.2021.126261](https://doi.org/10.1016/j.amc.2021.126261).

References

- [1] J. Aczél, F.S. Roberts, On the possible merging functions, *Math. Soc. Sci.* 17 (1989) 205–243.
- [2] A. Aldamak, S. Zolfaghari, Review of efficiency ranking methods in data envelopment analysis, *Measurement* 102 (2017) 161–172.
- [3] R. Allen, A. Athanassopoulos, R.D. Dyson, E. Thanassoulis, Weights restrictions and value judgments in data envelopment analysis: evolution, development and future directions, *Ann. Oper. Res.* 73 (1997) 13–34.
- [4] P. Andersen, N.C. Petersen, A procedure for ranking efficient units in data envelopment analysis, *Manag. Sci.* 39 (1993) 1261–1264.

- [5] T.R. Anderson, K. Hollingsworth, L. Inman, The fixed-weighting nature of a cross-evaluation model, *J. Prod. Anal.* 17 (2002) 249–255.
- [6] M.Z. Angiz, A. Mustafa, M.J. Kamali, Cross-ranking of decision making units in data envelopment analysis, *Appl. Math. Model.* 37 (2013) 398–405.
- [7] J. Aparicio, L. Ortiz, J.T. Pastor, J.M. Zabala-Iturriagagoitia, Introducing cross-productivity: a new approach for ranking productive units over time in data envelopment analysis, *Comput. Ind. Eng.* 144 (2020) 106456.
- [8] J. Aparicio, J.L. Zofio, Economic cross-efficiency, *Omega* 100 (2020) 102374.
- [9] J. Aparicio, J.L. Zofio, New definitions of economic cross-efficiency. Aparicio J., Lovell C.A.K., Pastor J.T., Zhu J. (eds), in: *Advances in Efficiency and Productivity II*, International Series in Operations Research & Management Science 281, (eds), Springer Science+Business Media, New York, 2020, pp. 11–32.
- [10] B.M. Balk, *Price and Quantity Index Numbers: Models for Measuring Aggregate Change and Difference*, Cambridge University Press, New York, 2008.
- [11] B.M. Balk, M.B.M. de Koster, C. Kaps, J.L. Zofio, An evaluation of cross-efficiency methods, applied to measuring warehouse performance, *ERIM Report Series in Management ERS-2017-015-LIS*, Erasmus University Rotterdam. The Netherlands, 2017.
- [12] J. Barbero, J.M. Zabala-Iturriagagoitia, Z.J. L., Benchmarking innovation systems with DEA-TOPSIS : on the relevance of decreasing returns on waning performance, *Forthcoming in Technovation*(2021).
- [13] I. Bardhan, W.F. Bowlin, W.W. Cooper, T. Sueyoshi, Models and measures for efficiency dominance in DEA part I: additive models and MED measures, *J. Oper. Res. Soc. Jpn.* 39 (1996) 322–332.
- [14] A. Charnes, W.W. Cooper, E. Rhodes, Measuring efficiency of decision making units, *Eur. J. Oper. Res.* 2 (1978) 429–444.
- [15] A. Charnes, W.W. Cooper, L. Seiford, J. Stutz, A multiplicative model for efficiency analysis, *Socio-econ. Plan. Sci.* 16 (1982) 223–224.
- [16] A. Charnes, W.W. Cooper, L. Seiford, J. Stutz, Invariant multiplicative efficiency and piecewise cobb-douglas envelopments, *Oper. Res. Lett.* 2 (1983) 101–103.
- [17] J.X. Chen, A comment on DEA efficiency assessment using ideal and anti-ideal decision making units, *Appl. Math. Comput.* 219 (2012) 583–591.
- [18] I. Contreras, Optimizing the rank position of the DMU as secondary goal in DEA cross-evaluation, *Appl. Math. Model.* 36 (2012) 2642–2648.
- [19] W.D. Cook, L.M. Seiford, Data envelopment analysis (DEA) – thirty years on, *Eur. J. Oper. Res.* 192 (2009) 1–17.
- [20] W.D. Cook, J. Zhu, Rank order data in DEA: a general framework, *Eur. J. Oper. Res.* 174 (2006) 1021–1038.
- [21] W.D. Cook, J. Zhu, DEA Cobb-Douglas frontier and cross efficiency, *J. Oper. Res. Soc.* 65 (2014) 265–268.
- [22] W.D. Cook, J. Zhu, DEA cross-efficiency, in: J. Zhu (Ed.), *Data Envelopment Analysis*, International Series in Operations Research & Management Science 221, Springer Science+Business Media, New York, 2015.
- [23] W.W. Cooper, J.L. Ruiz, I. Sirvent, Choices and uses of DEA weights. Cooper W.W., Seiford L.M., Zhu J. (eds), *Handbook On Data Envelopment Analysis*. New York, (eds.), Springer Science+Business Media, 2011.
- [24] M.B.M. De Koster, B.M. Balk, Benchmarking and monitoring international warehouse operations in europe, *Prod. Oper. Manag.* 17 (2008) 175–183.
- [25] M.B.M. De Koster, P.M.J. Warffemius, American, asian and third-party international warehouse operations in europe: a performance comparison, *Int. J. Oper. Prod. Manag.* 25 (8) (2005) 762–780.
- [26] J.R. Doyle, R.H. Green, Efficiency and cross-efficiency in DEA: derivations, meanings, and uses, *J. Oper. Res. Soc.* 45 (1994) 567–578.
- [27] J.R. Doyle, R.H. Green, Cross-evaluation in DEA: improving discrimination among DMUs, *Inf. Syst. Oper. Res.* 33 (1995) 205–222.
- [28] B. Ebrahimi, A. Dellnitz, A. Kleine, M. Tavana, A novel method for solving data envelopment analysis problems with weak ordinal data using robust measures, *Expert Syst. Appl.* 164 (2021) 113835.
- [29] Eurostat, Annual detailed enterprise statistics for services (NACRev. 2 h-n and s95), Luxemburg (2017). https://ec.europa.eu/eurostat/web/products-datasets/-/SBS_NA_1A_SE_R2.
- [30] N. Faber, M.B.M. de Koster, A. Smidts, Survival of the fittest: the impact of fit between warehouse management structure and warehouse context on warehouse performance, *Int. J. Prod. Res.* 56 (2018) 120–139.
- [31] M.J. Farrell, The measurement of productive efficiency, *J. R. Stat. Soc. Ser. A* 120 (1957) 253–281.
- [32] F.R. Forsund, Cross-efficiency: a critique, *Data Envelop. Anal. J.* 4 (2018) 1–25.
- [33] L. Friedman, Z. Sinyany-Stern, Scaling units via the canonical correlation analysis in the DEA context, *Eur. J. Oper. Res.* 100 (1997) 629–637.
- [34] J. Jablonsky, Measuring the efficiency of production units by AHP models, *Math. Comput. Model.* 46 (2007) 1091–1098.
- [35] G.R. Jahanshahloo, M. Khodabakhshi, F.H. Lotfi, M.M. Goudarzi, A cross-efficiency model based on super-efficiency for ranking units through the TOPSIS approach and its extension to the interval case, *Math. Comput. Model.* 53 (2011) 1946–1955.
- [36] A. Johnson, L.F. McInnis, Performance measurement in the warehousing industry, *IIE Trans.* 43 (2011) 220–230.
- [37] C. Kao, S.T. Liu, Cross efficiency measurement and decomposition in two basic network systems, *Omega* 83 (2019) 70–79.
- [38] F. Li, Q. Zhu, Z. Chen, H. Xue, A balanced data envelopment analysis cross-efficiency evaluation approach, *Expert Syst. Appl.* 106 (2018) 154–168.
- [39] L. Liang, J. Wu, W.D. Cook, J. Zhu, The DEA game cross efficiency model and its Nash equilibrium, *Oper. Res.* 56 (2008) 1278–1288.
- [40] S. Lim, J. Zhu, DEA cross-efficiency evaluation under variable returns to scale, *J. Oper. Res. Soc.* 66 (2015) 476–487.
- [41] J.S. Liu, L.Y. Lu, W.M. Lu, Research fronts in data envelopment analysis, *Omega* 58 (2016) 33–45.
- [42] W.-M. Lu, S.E. Lo, A benchmark-learning roadmap for regional sustainable development in China, *J. Oper. Res. Soc.* 58 (2007) 841–849.
- [43] R. Ma, L. Yao, M. Jin, P. Ren, The DEA game cross-efficiency model for supplier selection problem under competition, *Appl. Math.* 8 (2014) 811–818.
- [44] R. Maddahi, G.R. Jahanshahloo, F.H. Hosseinzadeh, A. Ebrahimnejad, Optimising proportional weights as a secondary goal in DEA cross-efficiency evaluation, *Int. J. Oper. Res.* 19 (2014) 234–245.
- [45] O.B. Olesen, Cross efficiency analysis and extended facets, *Data Envelop. Anal. J.* 4 (2018) 27–65.
- [46] H.H. Örkücü, V.S. Özsoy, M. Örkücü, H. Bal, A neutral cross efficiency approach for basic two stage production systems, *Expert Syst. Appl.* 125 (1) (2019) 333–344.
- [47] N. Ramón, J.L. Ruiz, I. Sirvent, On the choice of weights profiles in cross-efficiency evaluations, *Eur. J. Oper. Res.* 207 (2010) 1564–1572.
- [48] T.R. Sexton, R.H. Silkman, A.J. Hogan, Data envelopment analysis: Critique and extensions, in: R.H. Silkman (Ed.), *Measuring Efficiency: An Assessment of Data Envelopment Analysis, New Directions for Program Evaluation* 32, San Francisco/London: Jossey-Bass, 1986.
- [49] E. Thanassoulis, M.C.S. Portela, R. Allen, Incorporating value judgements in DEA, in: W. Cooper, L.W. Seiford, J. Zhu (Eds.), *Handbook on Data Envelopment Analysis*, Kluwer Academic Publishers, Boston, 2004.
- [50] C. Tofallis, On constructing a composite indicator with multiplicative aggregation and the avoidance of zero weights in DEA, *J. Oper. Res. Soc.* 65 (2014) 791–792.
- [51] A.M. Torgersen, F.R. Forsund, S.A.C. Kittelsen, Slack-adjusted efficiency measures and ranking of efficient units, *J. Prod. Anal.* 7 (1996) 379–398.
- [52] Y.M. Wang, K.S. Chin, Some alternative models for DEA cross-efficiency evaluation, *Int. J. Prod. Econ.* 128 (2010) 332–338.
- [53] Y.M. Wang, K.S. Chin, J.B. Yang, Measuring the performance of decision-making units using geometric average efficiency, *J. Oper. Res. Soc.* 58 (2007) 929–937.
- [54] J. Wu, J. Chu, J. Sun, Q. Zhu, DEA cross-efficiency evaluation based on Pareto improvement, *Eur. J. Oper. Res.* 248 (2) (2016) 571–579.
- [55] J. Wu, L. Liang, Y. Chen, DEA game cross-efficiency approach to olympic rankings, *Omega* 37 (2009) 909–918.
- [56] J. Wu, L. Liang, F. Yang, Determination of the weights for the ultimate cross-efficiency using shapley value in cooperative game, *Expert Syst. Appl.* 36 (2009) 872–876.
- [57] J. Wu, L. Liang, F. Yang, H. Yan, Bargaining game model in the evaluation of decision making units, *Expert Syst. Appl.* 36 (2009) 4357–4362.
- [58] J. Wu, J.S. Sun, L. Liang, Cross efficiency evaluation method based on weight-balanced data envelopment analysis model, *Comput. Ind. Eng.* 63 (2012) 513–519.
- [59] Y. Yamada, T. Matsui, M. Sugiyama, An inefficiency measurement method for management systems, *J. Oper. Res. Soc. Jpn.* 37 (1994) 158–168.
- [60] J. Zhu, *Quantitative Models for Performance Evaluation and Benchmarking*, Springer International Publishing Switzerland, 2014.
- [61] M. Zohrehbandian, S.S. Gavgani, Cross-efficiency evaluation under the principle of rank priority of DMUs, *World Appl. Sci. J.* 21 (2013) 46–49.