# An Evaluation of Cross-Efficiency Methods, Applied to Measuring Warehouse Performance

Bert M. Balk [*1], M.B.M. (René) De Koster [†1], Christian Kaps [‡1], and José L. Zofío [§2]

[1]Rotterdam School of Management, Erasmus University, Rotterdam, The Netherlands
[2]Department of Economics, Universidad Autónoma de Madrid, Madrid, Spain

Draft, December 6, 2017

**Abstract**

In this paper method and practice of cross-efficiency calculation is discussed. The main methods proposed in the literature are tested not on a set of artificial data but on a realistic sample of input-output data of European warehouses. The empirical results show the limited role which increasing automation investment and larger warehouse size have in increasing productive performance. The reason is the existence of decreasing returns to scale in the industry, resulting in suboptimal scales and inefficiencies, regardless of the operational performance of the facilities. From the methodological perspective, and based on a multidimensional metric which considers the capability of the various methods to rank warehouses, their ease of implementation, and their robustness to sensitivity analyses, we conclude to the superiority of the classic Sexton *et al.* (1986) method over recently proposed, more sophisticated methods.

**Keywords:** Cross-efficiency methods; warehouse efficiency; automation investment.

**JEL code:** C43, C61, D22, D24, L81, L87.

[*]E-mail: bbalk@rsm.nl.
[†]E-mail: rkoster@rsm.nl.
[‡]E-mail: Christian.Kaps@student.eur.nl.
[§]E-mail: jose.zofio@uam.es.

# 1 Introduction

Warehouses are undergoing profound changes from technological, operational and organizational perspectives, while their success in the market depends on whether they are capable of sustaining high levels of absolute and relative competitiveness. For management, competitive warehouse operations are key for market survival. Although the warehousing and storage industry in the EU accounted for 73 billion EUR in 2015 and the sector grows faster than the EU's GDP, there is a remarkable research gap in analyzing overall efficiency based on real-life data, making it arduous for businesses to answer these questions (Eurostat 2017).

Ample research has been dedicated to solving specific warehousing problems, which Gu *et al.* (2007) categorize into: 1) warehouse design, including structure, size, dimension, and department layout, and 2) warehouse operation, including receiving, shipping, storage, and order picking. However, the number of studies concerned with the aggregate efficiency of all warehouse interlinked input and output dimensions is limited. Indeed, many publications on how to improve single operational decisions have been published, yet besides their narrow focus, they often draw on simulation results (Van der Gaast 2015), small empirical samples (Larco 2016) or readily available data that had been collected for a different original purpose and consequently might not be as reliable as desired. Examples of relevant studies focusing on specific warehousing problems are the optimal number of zones in a pick-and-sort order picking system (De Koster *et al.* 2012) or the effect of picker personality on picking performance (De Vries *et al.* 2016). Their scope, however, is almost always limited to a single aspect of a warehouse.

One way to gauge how productive a given facility is, is to perform a comprehensive assessment of warehouse efficiency, which systematically reviews the warehouse functions with the intention of improving performance. One of the methods suited for multilateral comparisons of performance is Data Envelopment Analysis (DEA) because it provides a straightforward quantitative analysis of relative performance that ranks observations according to their relative productivity and, most importantly, quantitatively identifies real life peers that can be used as 'best-practice' references. Peers whose operational patterns and organizational characteristics can be mirrored so as to increase management achievements, including qualitative assessments of workload fulfillment processes, ways to manage inventory, role of automation and information technologies.

The strength of DEA is that it provides objective measurement, while identifying patterns and trends within the industry. It exactly portrays where firms stand and what they need to do in order to meet their goals. Moreover, in the event that a firm is inefficient, and therefore needs to adjust its production process, DEA results allow firms to focus on the various dimensions of the production process and to individualize the necessary changes that need to be undertaken. For example, it may be found that to improve performance with respect to its peers, a warehouse must make an efficient use of the available space and its personnel, whose value is in excess to its competitors producing the same level of services. In the same vein, a warehouse may have invested too heavily in automation, given its number of order lines, processing errors and level of order flexibility.

Two studies from the last decade have aimed at this goal, studying warehouse efficiency from a production perspective, based on a comprehensive input-output model. These are Johnson and McGinnis (2011) on warehousing performance in the United States, and De Koster and Balk (2008) on international distribution centers. Johnson and McGinnis (2011) collected a dataset for 216 warehouses between 2001 and 2005, using 5 outputs and 4 inputs which pass the 'efficiency contribution' test of Pastor *et al.* (2002). These authors find that warehouse efficiency is positively correlated with long term labor, high inventory turn, low seasonality, total replenishment, and better re-ordering practices. As a precursor of the present study, De Koster and Balk (2008) used survey methods to compile a database of 65 European distribution centers, producing 5 outputs from 4 inputs. These authors find mixed results as European distribution centers appear to be significantly more efficient that their North American and Asian owned counterparts (respectively 84%, 70%, and 72% in 2000), but experience a remarkable productivity decline from 2000 to 2004 of -12%, while their counterparts exhibit productivity growth rates of 3% and 6%, respectively.

However, though DEA is a powerful method to study productive performance, it is not without shortcomings when it comes to decision making. The most important weakness from a managerial perspective is the possibility of identifying processes which employ unrealistic quantities of inputs and outputs as efficient. This is because, in their multiplicative dual formulation, DEA methods search for the most favorable weights (shadow prices) when evaluating a production unit, frequently assigning zero values to some variables when constructing the 'virtual' aggregate input to output productivity ratio – each constructed as a linear combination of observed values.[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 when the optimal weights are inconsistent with prior knowledge of the production process, or when some inputs or outputs are ignored in the analysis.

A second consequence of the 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 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 as improbable weights are taken into account. This implies that the obtained ranking of firms misrepresents best-practice performance, as units whose 'virtual' production processes are either implausible from a managerial perspective, or infeasible from an engineering approach (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.

[1] Moreover, as remarked by Cooper *et al.* (2011), 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.

For example, the previous studies by Johnson and McGinnis (2011) and De Koster and Balk (2008) found that 23% and 45% of warehouses were efficient using the standard input oriented DEA approach. This makes it difficult to draw conclusions on warehouse best practices.[2]

To remedy these shortcomings there have been developed several proposals that restrict the construction of the virtual frontier, resulting in the identification of a credible set of efficient units, and thereby solving the discrimination problem. The first set of methods is characterized by the introduction of weight (shadow price) restrictions. Relative values within and between the input and output vectors are imposed in the multiplier form of the DEA problem – for a comprehensive exposition see Allen *et al.* (1997), Thanassoulis *et al.* (2004), and Cooper *et al.* (2011).[3]

A second set of methods uses the DEA version of the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS, Chen 2012). The technique creates virtual ideal (anti-ideal) production units with the maximum and minimum observed values of outputs and inputs, and calculates for each unit two efficiency scores, namely with respect to the ideal (optimistic) and anti-ideal (pessimistic) frontiers.[4]

The two previous approaches are capable of producing more sensible rankings by relying on reasonable weights and reducing their dispersion. The first modifies the production frontier by directly imposing weight restrictions, whereas the second does this indirectly, by constructing artificial (virtual) firms. Nevertheless, both approaches have as relevant drawback the fact that the optimal weights are not

[2]Under variable returns to scale the percentage of efficient warehouses increased to 69%. "In such case it does not make much sense to compare efficiencies anymore." (De Koster and Balk 2008, 181).

[3]These trade-off values link inputs and outputs according to some external value judgements based on prior information (e.g. market prices) and expert opinion (e.g. engineering shadow prices such as technical coefficients), and their relative importance helps to improve the discriminatory power of the method, while reducing weight dispersion. Setting lower and upper bounds for the weights results in the so-called assurance regions introduced by Thompson *et al.* (1986). More elaborated methods such as the analytic hierarchical process (AHP) were proposed by Saaty (1980). In a first step, the experts make pairwise comparisons between each pair of inputs (or outputs) and their judgement is quantified and recorded in a matrix. The AHP process transforms this information into a normalized weight vector in which individual elements represent the importance of each input (or output) in the process of evaluation.

[4]As DEA can be considered as an optimistic assessment of best practice in favor of the evaluated unit, resulting from the unrestricted choice of weights, including the ideal unit ensures dominance, and the problem of units being efficient by default is solved. Besides being ranked as inefficient, the farther away extreme units are from the ideal benchmark, the lower their efficiency value. Complementary is the formulation of a worst practice frontier through the introduction of the anti-ideal firm. Here the method seeks for the minimization of a unit's efficiency with respect to that frontier. The entire method consists of a two-stage process. In the first stage the efficiency scores of the ideal and anti-ideal units, excluding them from the production possibility set, are calculated (which effectively corresponds to the 'super-efficiency' approach of Andersen and Petersen 1993). In the second stage, an optimistic model (in the spirit of the benevolent approach in the following cross-efficiency methods) maximizes the relative efficiency of the evaluated unit under the condition that the best relative efficiency of the ideal unit remains unchanged. A pessimistic (or aggressive) model minimizes the relative efficiency of the unit while keeping the worst relative efficiency of the anti-ideal unchanged. Here extreme firms can be also penalized by conforming the worst practice frontier, Shen *et al.* (2016). The two scores are then combined into a composite performance index.

unique.

That there may exist an infinite number of solutions besides the one obtained by the simplex method in the software used 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 or transformation). Thus it is relevant to have some criterion for selecting a specific set of weights among all the optimal solutions. Ideally, such a criterion should also solve the ranking arbitrariness and provide a meaningful multilateral comparison of efficiencies among the units. As the ultimate objective of DEA is to provide comparisons of performance among the observed production units, the more bilateral evaluations that are brought into the analysis the more robust the rankings will be to partial (mis)representations of the production technology, as well as to extreme or unobserved firms (as the ideal and anti-ideal units).

Cross-efficiency is a method 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. Introduced by Sexton *et al.* (1986) and popularized by Doyle and Green (1994), cross-efficiency 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 so as 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 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.

These benevolent and aggressive secondary goals have been modified by other authors, as we will see in the methodological sections 2 and 3, including the multiplicative DEA approach by Cook and Zhu (2014). Wu *et al.* (2009a) and Liang *et al.* (2008) reinterpret the second-stage solution from the perspective of cooperative and non-cooperative games, respectively, while Wu *et al.* (2009b) exhaust this line of research by considering a bargaining game. Finally, a different strand corresponds to Wang and Chin (2010) and Ramón *et al.* (2010), where the weights for each unit are determined without considering the impact on its rivals.

In this article we explore the relative merits of the various cross-efficiency methods by assessing the operational performance of a set of 102 warehouses in the Netherlands and Belgium, in an attempt to answer specific questions related to the factors driving operational efficiency. The analysis looks into technological characteristics, differences among product categories, ownership types, and value chain positions, as well as recent changes in the industry between 2012 and 2017. This paper then contributes to both the methodological and the applied perspective.

The *methodological contribution* focuses on the relative merits of the various cross-efficiency methods. Based on a simple metric for ordered preference, we provide guidance when choosing among alternative methods. The methods are evaluated by the following criteria: (a) Ability to discriminate among warehouses, and to provide consistent and credible rankings for managerial decision making; that is, proximity

across methods in terms of statistical differences across the alternative rankings. (b) Extendability to real-life in-house business applications: implementational ease and computational requirements. (c) Sensitivity of the rankings to scale changes, erroneous entries, and removal of efficient peers from the reference frontier.

From an *empirical perspective* we are interested in testing the following hypotheses: (a) Do larger warehouses perform more efficiently? (b) Which input factors drive warehouse efficiency? (c) What are the effects of recent automation trends on warehouse efficiency? (d) How has relative warehouse efficiency changed over the last five years?

The paper unfolds as follows. In Sections 2 and 3 the cross-efficiency method is introduced, along with the most important secondary goals. In Section 4 the survey procedure used to collect the database is discussed, along with the representation of the warehouse technology by choice of variables included in the analysis. In Section 5 the calculations for the selected cross-efficiency methods are presented and the methodological and empirical questions are answered. Section 6 evaluates the performance of the methods according to the previous criteria. Section 7 concludes.

# 2 The basic idea of cross-efficiency measurement

Given input-output data for a set of firms (production units), for each observation a linear DEA program generates, next to an efficiency score, 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 an aggregate measure 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 is about how to select appropriate weights.[5]

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

Using DEA, for each firm $k = 1, ..., K$ its radial input technical efficiency (ITE), assuming constant returns to scale (CRS), is conventionally calculated as[6]

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

As is well known, $ITE(x^k, y^k)$, 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

[5] For an introduction to cross-efficiency see Cook and Zhu (2015) or Zhu (2014, Chapter 4).

[6] The restriction to the input orientation and CRS is for expository convenience. For the output orientation and variable returns to scale (VRS) one is referred to Cook and Zhu (2015). Under VRS input-orientated cross-efficiencies may become negative; see Lim and Zhu (2015) for a discussion of this problem. Notice that under CRS input technical efficiency and output technical efficiency are identical.

technical efficiencies are therefore used to compare firms. Put otherwise, firm $k$ is said to be more efficient than firm $\ell$ if $ITE(x^k, y^k) \geq ITE(x^\ell, y^\ell)$.

Is the use of $ITE(.)$ for comparing efficiencies warranted? To judge this we look at the dual LP problem[7]

$$\begin{aligned} &ITE(x^k, y^k) = \\ &\quad \max_{u,v}\{u \cdot y^k \mid u \cdot y^{k'} - v \cdot x^{k'} \leq 0, v \cdot x^k = 1, u \geq 0, v \geq 0, k' = 1, ..., K\} \\ &\quad = \max_{u,v}\left\{\frac{u \cdot y^k}{v \cdot x^k} \mid \frac{u \cdot y^{k'}}{v \cdot x^{k'}} \leq 1, u \geq 0, v \geq 0, k' = 1, ..., K\right\} \\ &\quad = \frac{u^k \cdot y^k}{v^k \cdot x^k}, \end{aligned} \tag{2}$$

where $(u^k, v^k)$ is a solution of the maximization problem. It now appears that

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

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, $(u^k, v^k)$ and $(u^\ell, v^\ell)$. These shadow prices are, in general, not unique. Let us, however, for the time being abstract from the nonuniqueness.

The situation depicted in expression (3) is a bit embarrassing. It would make more sense to base the efficiency comparison of the two firms on the comparison of either

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

where the weights of $k$ are used, or

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

where the weights of $\ell$ are used. This constitutes the idea behind the concept of cross-efficiency measurement.

Thus, the cross input technical efficiency (CITE) of firm $\ell$ with respect to firm $k$ is defined by

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

where $(u^k, v^k)$ satisfies equation (2). Notice that $CITE(x^\ell, y^\ell|\ell) = ITE(x^\ell, y^\ell)$ $(\ell = 1, ..., K)$. This could be called the self-appraisal score of firm $\ell$. The (arithmetic) mean appraisal score of firm $\ell$ by all its colleagues is given by

[7] This is known as the CCR model, after Charnes, Cooper and Rhodes (1978). 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.

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

The (arithmetic) mean overall appraisal score of firm $\ell$, called *the* cross input technical efficiency (CITE) of firm $\ell$, is then given by

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

which is a weighted mean of self-appraisal and colleague-appraisal scores, with weights $1/K$ and $(K-1)/K$, respectively.[8] Firm $\ell$ is now said to be more efficient than firm $\ell'$ 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

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

The interpretation of the measure defined by expression (5) in the single-input (or single-output) case was pointed out by Anderson, Hollingsworth and Inman (2002). When $N = 1$ the vectors $x$ and $v$ become scalars and their inner product reduces to simple multiplication. Then it appears that

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

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

We further notice that the scores of each appraiser $k = 1, ..., K$ constitute a (different) ratio scale, because each ratio $CITE(x^\ell, y^\ell|k)/CITE(x^{\ell'}, y^{\ell'}|k)$ for $\ell, \ell' = 1, ..., K$ admits a meaningful interpretation, namely as a productivity index (= ratio

[8] Lu and Lo (2007) show remarkable differences between self-appraisal and mean colleague-appraisal scores.

of output quantity index over input quantity index) of firm $\ell$ relative to $\ell'$. Then Aczél and Roberts (1989, Corollary 3.1) show that each ratio of merged scores

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

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

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

Then firm $\ell$ is more efficient than firm $\ell'$ 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.[9]

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

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

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.

# 3 The nonuniqueness problem

As noted below expression (2), the weight vectors $(u^k, v^k)$ $(k = 1, ..., K)$ are not unique. For instance, all extreme efficient units have an infinite number of optimal weights, as do all inefficient firms belonging to or projected onto the weak

[9] On Lowe price and quantity indices see Balk (2008). A Lowe-type productivity index is a ratio of Lowe output and input quantity indices.

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 $CITE(x^\ell, y^\ell|k)$ $(\ell, k = 1, ..., K)$ nor their means $\sum_{k=1}^{K} CITE(x^\ell, y^\ell|k)/K$ or $\prod_{k=1}^{K}(CITE(x^\ell, y^\ell|k))^{1/K}$ $(\ell = 1, ..., K)$ 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.

## 3.1 Aggressive and benevolent approaches

The idea behind the first approach is to select from the set of all the optimal weights solving the LP problem (2) those vectors which have the greatest discriminatory power. Based on Sexton, Silkman and Hogan (1986), Doyle and Green (DG) (1994, 1995) considered a number of options. The most natural is based on the $CITE(x^\ell, y^\ell|k)$ ratio formulation in expression (4):[10]

$$(u^{k*}, v^{k*}) \equiv \arg\min_{u^k, v^k} \left\{ \frac{1}{K-1} \sum_{k'=1, k' \neq k}^{K} CITE(x^{k'}, y^{k'}|k) \mid ITE(x^k, y^k) = \frac{u^k \cdot y^k}{v^k \cdot x^k} \right\}. \tag{8}$$

For each $k = 1, ..., K$ 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 (8) 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, but these days is not as we show in the empirical section.

As a feasible alternative Sexton, et al. (1986) had already considered[11]

$$(u^{k*}, v^{k*}) \equiv \arg\min_{u^k, v^k} \left\{ u^k \cdot \bar{y}^k - v^k \cdot \bar{x}^k \mid ITE(x^k, y^k) = \frac{u^k \cdot y^k}{v^k \cdot x^k} \right\}, \tag{9}$$

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 is 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

$$u^k \cdot \bar{y}^k - v^k \cdot \bar{x}^k = v^k \cdot \bar{x}^k \left( CITE(\bar{x}^k, \bar{y}^k|k) - 1 \right).$$

To overcome this problem Doyle and Green (1994, 1995) considered[12]

[10] This is option I of DG (1994) = option III of DG (1995). Notice that the options are differently numbered in the two papers.

[11] This is option II of DG (1994) = option I of DG (1995).

[12] This is option III of DG (1994) = option II of DG (1995).

$$(u^{k*}, v^{k*}) \equiv \arg\min_{u^k, v^k} \left\{ CITE(\bar{x}^k, \bar{y}^k|k) \mid ITE(x^k, y^k) = \frac{u^k \cdot y^k}{v^k \cdot x^k} \right\}. \tag{10}$$

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 (8) is replaced by the ratio of means in problem (10).

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

## 3.2 The multiplicative variant

Charnes *et al.* (1982) introduced the multiplicative variant of the LP problem (2):

$$\begin{aligned} &ITE'(x^k, y^k) \equiv \\ &\max_{u,v} \left\{ \frac{\prod_{m=1}^{M} (y_m^k)^{u_m}}{\prod_{n=1}^{N} (x_n^k)^{v_n}} \mid \frac{\prod_{m=1}^{M} (y_m^{k'})^{u_m}}{\prod_{n=1}^{N} (x_n^{k'})^{v_n}} \leq 1, u \geq 1, v \geq 1, k' = 1, ..., K \right\} \\ &= \frac{\prod_{m=1}^{M} (y_m^k)^{u_m^k}}{\prod_{n=1}^{N} (x_n^k)^{v_n^k}}, \end{aligned} \tag{11}$$

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

$$CITE'(x^\ell, y^\ell|k) \equiv \frac{\prod_{m=1}^{M} (y_m^\ell)^{u_m^k}}{\prod_{n=1}^{N} (x_n^\ell)^{v_n^k}} \ (\ell, k = 1, ..., K), \tag{12}$$

where $(u^k, v^k)$ satisfies equation (11). Cook and Zhu (2014), (2015) proposed to merge these scores by a geometric mean, and defined

$$\max_{u^k, v^k} \left\{ \prod_{k=1}^{K} (CITE'(x^\ell, y^\ell|k))^{1/K} \mid ITE'(x^k, y^k) = \frac{\prod_{m=1}^{M} (y_m^k)^{u_m^k}}{\prod_{n=1}^{N} (x_n^k)^{v_n^k}}, k = 1, ..., K \right\} \tag{13}$$

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

It might be seen as a problem that the function defined by expression (11) is not invariant to changes in the units of measurement of the inputs and outputs. The solution provided by Charnes *et al.* (1983) is to consider a slight modification of the foregoing maximization problem, namely by inserting a scalar $\omega$ so that

$$\begin{aligned} &ITE''(x^k, y^k) \equiv \\ &\max_{u,v,\omega} \left\{ e^\omega \frac{\prod_{m=1}^{M} (y_m^k)^{u_m}}{\prod_{n=1}^{N} (x_n^k)^{v_n}} \mid e^\omega \frac{\prod_{m=1}^{M} (y_m^{k'})^{u_m}}{\prod_{n=1}^{N} (x_n^{k'})^{v_n}} \leq 1, u \geq 1, v \geq 1, k' = 1, ..., K \right\} \end{aligned} \tag{14}$$

$$= e^{\omega^k} \frac{\prod_{m=1}^{M} (y_m^k)^{u_m^k}}{\prod_{n=1}^{N} (x_n^k)^{v_n^k}},$$

where $(u^k, v^k, \omega^k)$ is a solution of the maximization problem. By taking logarithms, this appears to be a VRS additive DEA model. Tofallis (2014) points out an additional benefit of this model: there is "no longer any fear of zero (or infinitesimal epsilon) weights."

### 3.3 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.* (2008) took another route. These authors considered the following modification of the LP problem (2):

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

$$= \frac{u^k(\alpha^{dt}) \cdot y^k}{v^k(\alpha^{dt}) \cdot x^k}, \tag{15}$$

where the shadow price vectors $(u^k(\alpha^{dt}), v^k(\alpha^{dt}))$ solve the maximization problem, $d = 1, ..., K$, 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 $\alpha^{dt}$. Obviously, all the weights are then functions of this level. Notice that the constraints imply that $\alpha^{dt} \leq 1$ $(d = 1, ..., K)$.

If the vector pair $(u^k(\alpha^{dt}), v^k(\alpha^{dt}))$ solves the maximization problem (15) then

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

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 (15) implies that $\alpha^{dt} \leq ITE(x^d, y^d)$.

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

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

This, however, means that $(u^d, v^d)$ satisfies the conditions defining the maximization problem (15), and hence

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

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

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

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

$$\alpha^{k,t+1} \equiv \frac{1}{K}\sum_{d=1}^{K}\frac{u^k(\alpha^{dt})\cdot y^k}{v^k(\alpha^{dt})\cdot x^k}\ (k=1,...,K); \tag{16}$$

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 $\alpha^{dt}$.

Basically, expression (16) defines a mapping from the set $[0,1]^K$ to $[0,1]^K$. For $\alpha^{dt} \in [\sum_{k=1}^{K} CITE(x^d, y^d|k)/K, ITE(x^d, y^d)]$ $(d=1,...,K)$ Liang *et al.* (2008) showed that this mapping is continuous. Thus, by Brouwer's Fixed Point Theorem, there exists a vector of $\alpha$'s such that

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

Liang *et al.* (2008) also showed that the iterative system built on expression (16) converges.

Initially, the levels are thereby chosen as

$$\alpha^{\ell 0} = \sum_{k=1}^{K} CITE(x^\ell, y^\ell|k)/K\ (\ell=1,...,K); \tag{17}$$

that is, the mean cross input technical efficiencies generated by the original DEA problems (2). Then the levels $\alpha^{\ell 1}$ $(\ell=1,...,K)$ are generated by applying expression (16), *etcetera*, until convergence is reached.

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.* (2008). Wu *et al.* (2009a) developed an alternative model from the perspective of cooperative games, while Wu *et al.* (2009b) went further by using the perspective of a bargaining game.

# 4 Implementation and research hypotheses

## 4.1 Warehouse processes and variable selection

As anticipated in the Introduction, for DEA models to have sufficient discriminatory power and provide meaningful rankings, large degrees of freedom are necessary, with the balance between observations and variables playing a pivotal role.

As De Koster and Warffemius (2005), De Koster and Balk (2008), and Faber *et al.* (2013) extensively surveyed and analyzed warehouses, we rely on their choice of

input and output variables to characterize the warehouse processes.[13] This choice is qualified by the new research questions we address in the present study, in particular those related to the effect that increasing investments in new hardware and software automation processes, coupled with larger warehouse sizes, have on efficiency and cross-efficiency evaluations.

Following Johnson and McGinnis (2011), we apply newly available tests aimed at determining whether there exists any redundancy, and perform the specification test proposed by Pastor *et al.* (2002), based on the so-called 'efficiency contribution measure'. While De Koster and Balk (2008) favored the use of four inputs and five outputs, Faber *et al.* (2013) reduced the output dimension by combining the formerly separate outputs of *value-added logistics* and *special processes* into one construct, as in the previous study value-added logistics by itself did not significantly impact efficiency.

On these grounds, and referring the reader to the above mentioned publications for a discussion of the merits of these variables and constructs and their operational impact, our preferred DEA model consists of the following input and outputs variables collected in 2012 and 2017.

### 4.1.1 Four input factors

1. *Warehouse size in* $m^2$ (Floor space): Measured on a ratio scale, capturing the floor space of the warehouse, including mezzanine floor.

2. *Number of Full Time Equivalent employees* (FTEs): Measured on a ratio scale, including employees and temporary workers (1 fte = 1700 hours).

3. *Number of Stock Keeping Units* (SKUs): Measured on a ratio scale as the average number of unique articles that are simultaneously stored in the warehouse, in 2012 and 2017 each.

4. *Level of automation* (Automation): Measured on an ordinal scale, as the sum of an ordinal score for hardware automation and an ordinal score for software automation. Hardware automation is calculated based on how many out of 14 common automation technologies are employed in a warehouse. For the temporal comparison, each warehouse also answered which automation additions it had made during the last 5 years. The 2017 automation score is then calculated as the sum of the 2012 score plus all additions. Next to hardware automation, software automation is measured on a six-point ordinal scale, through a question about the warehouse's usage of information systems.[14]

### 4.1.2 Four output factors

1. *Number of order lines* (Order lines): Measured on a ratio scale. Refers to average order lines shipped per day during 2012 and 2017, respectively.

[13] Hackman *et al.* (2001) Johnson and McGinnis (2011) used similar variables.

[14] The possible answers were: (1) No information system; (2) a standard ERP warehouse module; (3) a standard ERP warehouse module with more than 20% customization; (4) a standard WMS package; (5) a standard WMS package with more than 20% customization, or (6) a tailor-made/customized system.

2. *Error-free order line percentage*(Error free %): Measured on an nine-point ordinal scale, with the following values: (1) Not tracked (2) <90%; (3) 90-95%; (4) 95-97%; (5) 97-98%; (6) 98-99%; (7) 99.0-99.5%; (8) 99.5-99.9% and (9) >99.9%. Not tracking this metric (or not having the data available) is penalized, as the error-free percentage is a very important quality criterion in warehousing and not observing it renders most internal quality control efforts irrelevant. By providing staggered levels of error-free order lines, this model can differentiate between a wide array of error-free percentages.

3. *Order flexibility*: Measured on a 30-point ordinal scale. Each respondent was asked whether his/her warehouse could cope with a total of six internal and external changes (1) much worse; (2) worse; (3) equal; (4) better or (5) much better than its competition.[15]

4. *Special processes*: Measured on a 10-point ordinal scale, where respondents selected from a list of ten special value-added processes possibly be executed by the warehouse. The number of selections in each year was used as the score.

## 4.2 Research hypotheses

Based on the previous characterization of warehouse production structure and organization, and recent trends related to automation levels, production scale, and (product) specialization, the research hypotheses addressed are the following.

- *Hypothesis 1a and 1b*: (Automation intensity in levels and rates of change, respectively): Despite the characterization of automation as productive input, its correlation with warehouse efficiency is expected to be positive. As warehouses are investing heavily in both IT hardware and software, this should be reflected in more automated facilities exhibiting higher performance levels. Instead of levels hypothesis 1b considers rates of change between 2012 and 2017 as it is generally acknowledged that the benefits of investing in IT may be realized after a temporal lag.

Hypotheses 1a and 1b provide an efficiency explanation to the observed increase of warehouse automation investments in recent years, which is being conceived as a change in paradigm by authors such as Wang *et al.* (2010). A similar proposition was made by Hamberg and Verriet (2012), who believed that as the technical hurdles for some of the most complex problems in warehousing would be overcome (by unstructured automatic item picking, autonomous vehicle roaming), a transformation toward automated warehousing would likely occur.

- *Hypothesis 2 and 3* (Size in SKUs and FTEs, respectively): Based on previous findings in the literature, size measured from the input perspective (by SKUs or FTEs) and productive efficiency are negatively correlated. As there has

[15]When "not applicable" had been selected, the respective question was taken out of consideration for that warehouse and its score re-scaled. Re-scaling was achieved by eliminating the "not applicable" questions from a DMU's score and multiplying the remaining score by 6 divided by the number of questions that were applicable.

been an increase in the average number of SKUs and FTEs in the industry over the last years, we presume that there might be a reversal in this relationship as facilities are expanding their scale of operations. Also, given the importance of employment from a social standpoint, it is relevant to explore its relationship (complementarity or substitutability) with other inputs like capital (floor space) and automation.

Size related input factors such as the number of SKUs, the number of FTEs, or floor space play a role in performance assessment related to the (sub-) optimality of the chosen production scales, and a warehouse's ability to attain the most productive scale size. As Hackman *et al.* (2001) and De Koster and Balk (2008) independently found small warehouses (measured in FTEs) to be more efficient, we hypothesize that there might be a reverse relationship.

Moreover, extending these and previous contributions, we decompose technical efficiency into operational (VRS) and scale efficiency. Facilities may grow in size to accommodate increasing demand, but the industry may contemporarily exhibit decreasing returns to scale. Then, as long as scale inefficiency is kept to a minimum, managers cannot be held responsible for lower productivity levels; i.e., those corresponding to the efficiency scores under constant returns to scale.[16] This technological qualification may help understand why firms undertake plant extensions – or directly build larger facilities from scratch, even if the technology is subject to decreasing returns to scale. In the case of decreasing returns, large individual warehouses serving as hubs are penalized.

As there is a constant shift towards facilities with larger footprint (floor space), it is legitimate to conjecture that these are designed to reap the (suspected) benefits of economies of scale, as not all size-related input factors are likely correlating negatively with efficiency. A report from Onstein *et al.* (2016) found that especially in the Netherlands "the growing demand for very large DCs" is a dominant phenomenon, most prominently for e-commerce logistics activities.

- *Hypothesis 4a and 4b* (Floor space ($m^2$) in levels and rates of change, respectively): It is expected that floor space is positively correlated with both scale and operational (VRS) efficiencies, situating the largest facilities at the top of the cross-efficiency rankings, in accordance with a technology characterized by non-decreasing returns to scale. Also, from a temporal perspective, and driven by the increase in floor space size, one may assume that the relative change in the available floor space correlates positively with the relative change in the warehouses' productive efficiency and cross-efficiency rank.

Figure 1 shows how the various hypotheses are related.

[16]It is worth noting that the hypothesis is not built on the general assumption of decreasing returns of scale, as a newly added SKU is not inherently less efficiently handled than an existing one. Also, two FTEs may pick twice the number of order lines of one FTE (as would be the case when constant returns to scale prevails). The assumption of lower efficiency for higher SKUs is based on the fact that more SKUs may make it more difficult to have standardized processes, or to keep high turnovers of products, and will lower workers' familiarity with each product. Likewise, more FTEs may lead to a lower degree of identification with the company, lower social pressure, more formalized rules and less specific talent being hired.

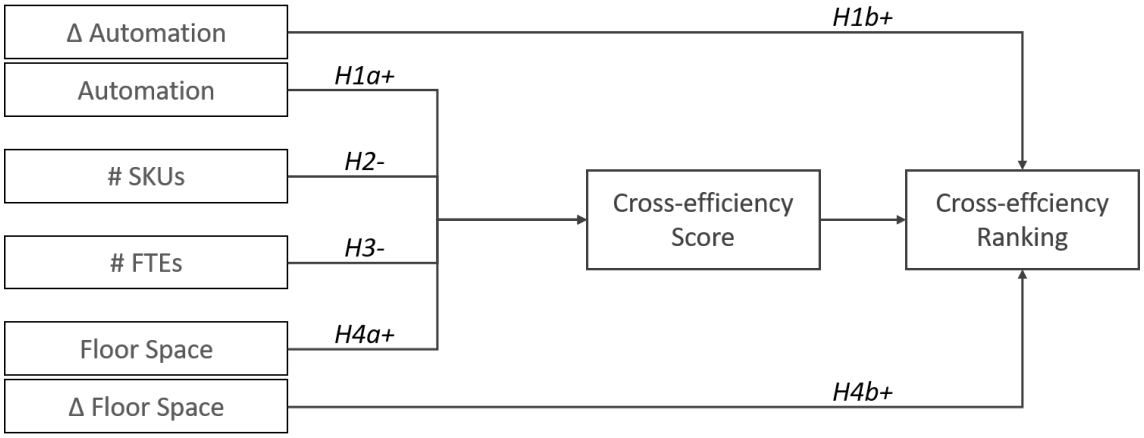


Figure 1: The hypotheses.

## 4.3 Survey methods and data

The database used in the study is the result of a comprehensive on-line reach-out- and response collection process. The mailing lists of two warehouse associations, evofenedex and TLN, which collaborated in the reach-out process and supported the study as interested stakeholders, and a large third company with multiple distribution centers, as well as contact data of previous research in the warehousing sector by the authors, were used to gather respondents. Additionally, extensive personal outreach via Linkedin, RSM's supply chain alumni network, and online search was conducted. Each contacted individual was introduced to the subject of the study and incentivized by offering an individual report of their facility's performance compared to the competition.[17] To define the questionnaire, several drafts were pilot-tested by a small group of logistics professionals; they commented on the questions, perceived ambiguities, and assessed the time needed to complete it.[18] The final questionnaires were administered in March-May 2017 through *Qualtrics*, an online survey platform, where the survey was accessible in Dutch and English. Through the browser-based interface respondents entered the data related to the warehouse inputs and outputs, as well as their attributes: product category, value chain position, ownership, *etcetera*.

All in all, 1,827 individuals (warehouse, logistics, and supply chain managers) were contacted, which resulted in 214 submissions, of which 131 were completely filled out and useful.[19] The response rates are 11.7% (all respondents) and 7.2% (useful respondents), which is lower than the response rates Muilerman (2001) identified for studies in the logistics sector. However, that study was published before the ad-

[17]The online material consisting of the introductory letter, questionnaire, and sample report are available upon request and will be accessible as online appendices.

[18]The supply-chain-focused websites logistiek.nl, logistiekProfs.nl and logistiektotaal.nl each published articles about the study. The total view count however cannot be established.

[19]The three most often cited reasons to decline were that managers are 1) too busy, 2) not allowed to share the required data, 3) not interested in external benchmarking. The first reason accounted for over 90%.

vent of the internet and is therefore not directly comparable to online-administered surveys as in this study. Sauermann and Roach (2013) find in their study that the low costs of surveying with online tools have led to "oversurveying", thereby reducing the achievable response rates below earlier levels.

### 4.3.1 Final data set and clusters distinguished

Of the 131 complete surveys, 102 were used for the analysis. The remaining 29 surveys were eliminated because these warehouses employed less than 5 FTEs, making them extreme observations and therefore hardly comparable to full-size warehouses.[20] By nationality, 82 warehouses are located in the Netherlands, 17 in Belgium and 3 in Germany. The warehouses store 12 different product categories (all 13 selectable categories, except Military/defense), employ over 6,000 FTEs and store 2.2 million SKUs on over 1.8 million square meters, of which 25% are Cold-storage. The initial data set of respondents contains warehouses from all industries, all positions in the value-chain and all ownership types.

By product category, General logistics accounts for 26% of the respondents, followed by Consumer goods", 20%, Food and groceries, 19%, and Retail (non food), 14%.[21] This diversity has varying effects on the comparability of warehouses, and shows the need for clustering warehouses into sets capable of providing insights on efficiency patterns across sectors. Because of the limited number of observations per industry, no universal applicability for these analyses can be claimed. Nevertheless, this drill-down and the intermediate juxtaposition of homogeneous warehouses was the most requested feature by participating managers. Taking into consideration the recommendations for the minimum number of observations in a DEA analysis, only clusters formed with approximately 20 or more observations provide meaningful efficiency analyses and cross-efficiency rankings. This strikes a balance between sufficient discrimination of units and homogeneity of observations when performing sectoral analyses.

On these grounds, the following three clusters representing a combination of two product categories with comparable operations were defined: 1) Engineering plus Construction (21 warehouses); 2) Consumer goods (20 warehouses); and 3) Groceries and Food (19 warehouses). For each cluster of warehouses we perform a separate cross-efficiency analysis. We expect that the dispersion of cross-efficiencies in the entire set is larger than in the clusters, where competitive pressure plays a role.

[20] The previously discussed discriminatory power of DEA is empirically addressed in several guidelines offered in the literature about the required minimum number of observations (Avkiran 2006). Our final number of warehouses ensures that the dataset fulfills the aggressive rule-of-thumb of Dyson *et al.* (2001) – the minimum number of observations must be equal to the number of inputs multiplied by the number of outputs times two, or 32 in our case. Golany and Roll (1989) only recommend twice the total number of input and output variables, or 16 in our case. This issue will be revisited in the discussion of clusters.

[21] Each respondent was allowed to select up to two product categories that best classified the warehouse's products. The 102 warehouses selected 145 categories, with General logistics, Consumer goods and Groceries and food stated most often, classifying half of the warehouses. Although General logistics exhibits the highest category share, it is also the category which was most often co-selected together with another product category.

With respect to ownership, 57% of the warehouses are owned and operated In-house, 33% is operated by a Third party logistics provider (3PL) in a warehouse facility offering services to multiple customers, and only 10% of the warehouses are dedicated to a single company and operated through a provider. 60% of the warehouses are Wholesale warehouses, shipping to B2B partners, and 26% are Production warehouses. Facilities are considered production warehouses, if the production of the products stored happens on the same premises. This includes raw materials and components as well as finished products for further distribution. 14% of the warehouses are Retail facilities with direct end-customer contact. It is assumed that a warehouse's position within the value chain has a large impact on the achievable efficiency, see Emmett (2005).

For a matrix overview of the two dimensions, see Table 1.

Table 1: Value chain position and operations provider of the warehouses in the sample.

| | In-House | 3PL-Dedicated | 3PL-Multiple | Sum |
|---|---|---|---|---|
| Production | 19 | 0 | 8 | 27 |
| Wholesale | 31 | 6 | 24 | 61 |
| Retail | 8 | 4 | 2 | 14 |
| Sum | 58 | 10 | 34 | 102 |

### 4.3.2 Size breakdowns and returns to scale

To study the nature of returns to scale and the underlying scale efficiency performance it is necessary to have a picture of the dataset in terms of size distribution. Size can be measured from the input or the output perspective. As shown by Banker and Thrall (1992), however, both perspectives are involved as returns to scale is determined by the sum of the multipliers in the primal DEA program under constant returns to scale, or the sign of the scale parameter in the dual multiplier formulation.

A graphical representation of two size distributions can be found in Figures 2 and 3. In both cases the mean size has grown over the past 5 years. For SKUs and order lines, the relative differences in size became larger across the sample, which is explainable through their different nature. For instance, an aircraft spare-part warehouse, maintaining stock of all parts over 40 years, versus a fresh-fruit importer, where no item is stored for longer than 24 hours. The SKU range moved from 100-250,000 in 2012 to 100-400,000 in 2017, and the range of order lines from 25-55,000 in 2012 to 54-55,000 in 2017. Also, the range of FTEs increased from 5-312 in 2012 to 5-350 in 2017, and the range of floor space (in $m^2$) from 400-275,000 in 2012 to 500-275,000 in 2017.

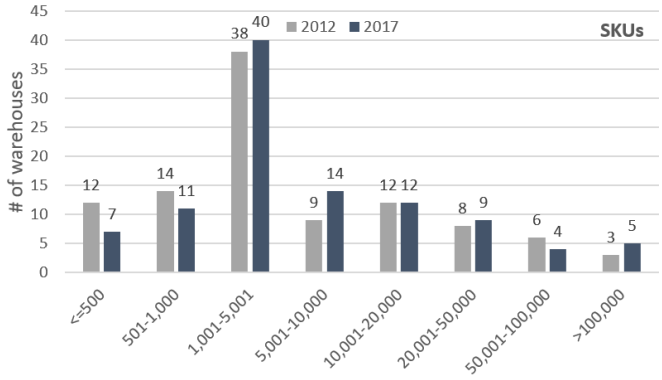


Figure 2: Number of SKUs across the warehouse sample in 2012 and 2017.

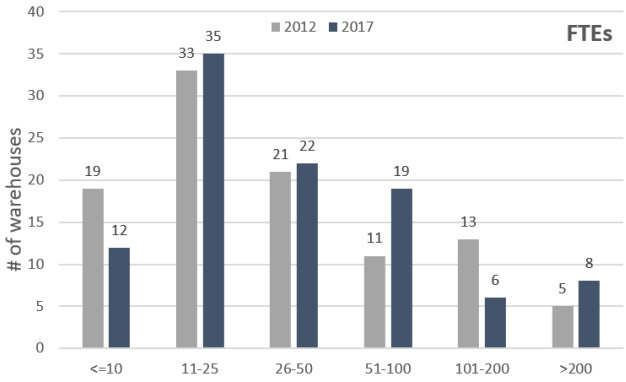


Figure 3: Number of FTEs across the warehouse sample in 2012 and 2017.

### 4.3.3 Trends and relevance

Tables 2 and 3 present the descriptive statistics for input and outputs in 2017, respectively. In the last row of both tables the percentage change from 2012 to 2017 is shown. There is a clear increase in the scale of operations as the average number of SKUs has grown by more than 10%, floor space and automation by almost 40%. These trends show that an increase in size goes hand-in-hand with both hardware and software improvements, and the substitution of labor by capital (e.g., robotics).

The increase in size is also observed at the output side. To the extent that on average output quantities have increased more than input quantities it can be concluded that the industry has increased its productivity. Notice that order lines as well as the percentage of error free order lines have increased by almost 40%. We return to these size changes when determining the nature of returns to scale and scale efficiency.

Table 2: Descriptive statistics of inputs, 2017.

| | FTEs | Floor space ($m^2$) | SKUs | Automation score |
|---|---|---|---|---|
| Minimum | 5 | 500 | 100 | 2 |
| Median | 30 | 9,250 | 4,600 | 6 |
| Average | 59 | 18,244 | 21,088 | 7 |
| Maximum | 350 | 275,000 | 400,000 | 16 |
| Standard dev. | 74 | 32,414 | 57,393 | 3 |
| $\Delta$(%) 2017/12 | 14.2 | 37.2 | 11.8 | 38.1 |

Table 3: Descriptive statistics of outputs, 2017.

| | Order lines | Special process | Error free % | Order flexibility |
|---|---|---|---|---|
| Minimum | 54 | 2 | 1 | 12 |
| Median | 1,200 | 6 | 7 | 22 |
| Average | 4,931 | 6 | 6 | 21 |
| Maximum | 55,000 | 10 | 9 | 30 |
| Standard dev. | 9,815 | 2 | 2 | 4 |
| $\Delta$(%) 2017/12 | 34.2 | 24.5 | 39.6 | 18.2 |

The relevance of all the input and output variables has been tested by the efficiency contribution measure (ECM) of Pastor *et al.* (2002). Departing from the solutions to DEA program (2), calculated with all the inputs and outputs, a second round of technical efficiency scores is calculated by dropping one input or output at a time. If these scores shrink less than 10% for 90% of the observations, then the information provided by the removed variable is redundant, and can safely be deleted from the model. The iterative test stops when no additional variable can be removed without losing a significant amount of information. All the variables included in the model passed the ECM test and none had to be removed.[22]

# 5 Calculations

This section first explores the technological characteristics of warehouse operations so as to determine the role that returns to scale play. Thereafter the results of the various cross-efficiency methods will be discussed.

## 5.1 Returns to scale and scale efficiency

### 5.1.1 Identifying returns to scale

By applying DEA model (2), 35 warehouses appear to be technically efficient in 2012 and 26 in 2017. The drop in the number of efficient firms is linked to a decrease in

[22]Detailed test results are available upon request.

the industry's average efficiency scores from 0.751 in 2012 to 0.681 in 2017 (with a standard deviation of 0.25 in both cases). Of the 26 efficient warehouses in 2017, 11 act as peers for 10 or more rivals (e.g., WH#73 for 51 inefficient counterparts which is almost half of the sample, WH#33 for 48, and WH#20 for 37), while others can be considered as extreme or "maverick" observations since they do not constitute a reference for any counterpart (e.g., WH#6 or WH#25). This lets one infer that the efficient operations most often selected as peers are more representative for likely efficiency improvement strategies in the form of input and output changes. Indeed, any firm not efficient under constant returns to scale (CRS) is scale inefficient, the nature and severity of which can be discerned by

(i) studying the nature of returns to scale, increasing (IRS) or decreasing (DRS), at their benchmark frontier; and

(ii) solving the variable-returns-to-scale (VRS) counterpart to program (1), followed by their scale efficiency value defined as the ratio of DEA-CRS to DEA-VRS scores.

As to (i), it should be understood that the concept of returns to scale is unambiguous only at frontier points. Following Banker and Thrall (1992), we start from solving the primal envelopment form (1). For a certain firm $k$ let $ITE(x^k, y^k) = \delta^{k*}$ and let the associated lambdas be $\lambda^*_{k'}$ $(k' = 1, ..., K)$. Positive values indicate peers for firm $k$. The (radial) projection of firm $k$ on the frontier is given by $(\sum_{k'=1}^{K} \lambda^*_{k'} x^{k'}, \sum_{k'=1}^{K} \lambda^*_{k'} y^{k'})$.

The returns to scale at this point on the production frontier can then be determined in the following way:[23]

(i) If $\sum_{k'=1}^{K} \lambda^*_{k'} = 1$ for at least one optimal solution, then CRS prevails.

(ii) If $\sum_{k'=1}^{K} \lambda^*_{k'} > 1$ for all optimal solutions, then DRS prevails.

(iii) If $\sum_{k'=1}^{K} \lambda^*_{k'} < 1$ for all optimal solutions, then IRS prevails.

Fortunately, there is no need to go through all the solutions, as shown by Banker, Chang and Cooper (1996). Suppose that we obtained $\sum_{k'=1}^{K} \lambda^*_{k'} > 1$. The next program then checks whether there exist alternative solutions compatible with the same efficiency level, but where the sum of lambdas equals 1 (and CRS, rather than DRS, prevails as in (i) above). Notice that the objective function adjusts the sum of lambdas by accounting for eventual slacks $(s^-, s^+)$ through the infinitesimal weight $\epsilon$. Thus, the solution with the largest total slack is obtained.

$$\min_{\lambda, s^-, s^+} \{ \sum_{k'=1}^{K} \lambda_{k'} - \epsilon \left( \sum_{n=1}^{N} s_n^- + \sum_{m=1}^{M} s_m^+ \right) \mid \sum_{k'=1}^{K} \lambda_{k'} x^{k'} + s^- = \delta^{k*} x^k, \tag{18}$$

$$y^k + s^+ = \sum_{k'=1}^{K} \lambda_{k'} y^{k'}, \sum_{k=1}^{K} \lambda_{k'} \geq 1, \lambda_{k'} \geq 0, k' = 1, ..., K \}.$$

It is clear that the original solution satisfies the conditions of the last problem. Thus, if the optimal value of the objective function appears to be larger than 1 then DRS prevails.

[23] It is also possible to identify the nature of returns to scale from the VRS specification of the dual multiplier form (2); that is, the BCC model, after Banker, Charnes and Cooper (1984).

If at the first stage we had obtained $\sum_{k'=1}^{K} \lambda_{k'}^{*} < 1$, then minimization must be replaced by maximization, the objective function in expression (18) must be adjusted to

$$\sum_{k'=1}^{K} \lambda_{k'} + \epsilon \left( \sum_{n=1}^{N} s_n^{+} + \sum_{m=1}^{M} s_m^{-} \right), \tag{19}$$

and the sum of lambdas must be constrained to be smaller than or equal to 1. It is clear that Pareto-efficient firms ($\delta^{k*} = 1$, $(s^{-}, s^{+}) = (0, 0)$) exhibit local CRS.

Applying this analysis to our database delivers the results summarized in Table 4. Interestingly the more recent year exhibits a higher number of facilities operating in the DRS region. To convey an idea of the returns to scale in the sample, Table 5 shows the operationally most efficient range of inputs for warehouses; that is, the range *below* which only warehouses with IRS were found and *above* which only warehouses with DRS were found. This helps to understand at which scale of inputs, facilities tend to be most efficient, disregarding the operational strategies they employ.

With a wide variety of processes it is possible to achieve optimal scale size, yet there are thresholds on floor space, assortment size as well as the level of automation. For FTEs, the smallest and largest facilities were nearly all within the optimal range.

Table 4: Warehouse returns to scale.

| | 2012 | 2017 |
|---|---|---|
| IRS | 25 | 22 |
| DRS | 42 | 54 |
| CRS | 35 | 26 |
| Total | 102 | 102 |

Table 5: Range of optimal input scales for 2012 and 2017.

| | 2012 | | 2017 | |
|---|---|---|---|---|
| | Min | Max | Min | Max |
| FTEs | 8 | 115 | 10 | 243 |
| Floor space | 1,500 | 26,000 | 1,260 | 275,000 |
| SKUs | 700 | 250,000 | 850 | 17,600 |
| Automation | 3 | 10 | 4 | 14 |

### 5.1.2 Decreasing returns to scale for large warehouses

It is relevant to investigate what type of returns to scale prevail the data set by warehouse size, and whether size increasing trends can explain the observed shift in RTS over time. Previously, studies found large warehouses to be less efficient than small warehouses (De Koster and Balk, 2008), but it is critical to distinguish between two factors that may drive this efficiency gap. First, DRS could be at play,

which would make it impossible for larger warehouses to become efficient by simply reducing operational – technical – inefficiency, as achieving an optimal scale would involve changing the relative size of the facility. Second, operational inefficiencies could be the reason that large warehouses are less efficient.

A visual inspection of returns to scale obtained through Banker and Thrall's (1992) method for alternative magnitudes of FTEs, floor space and SKUs, as plotted in Figure 4, clearly shows decreasing returns for larger warehouses.

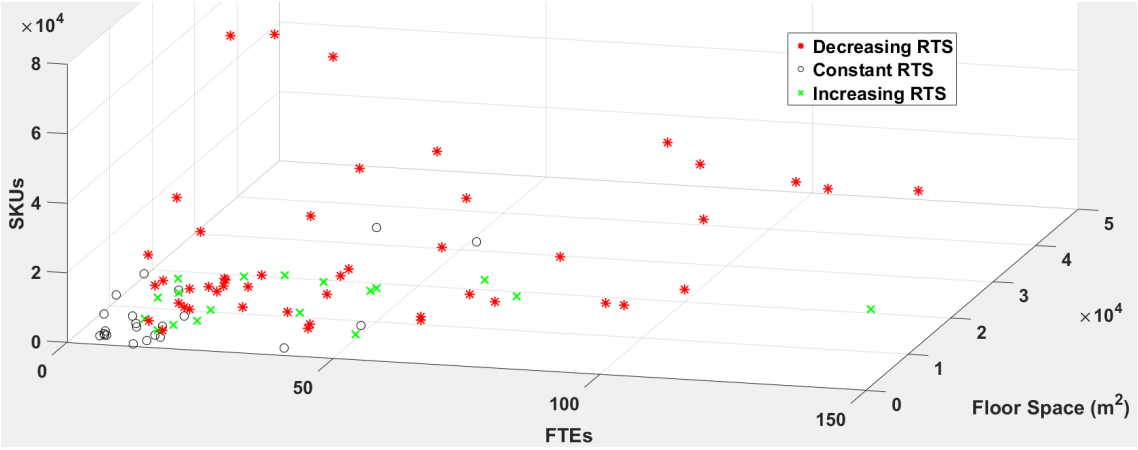


Figure 4: Returns to scale compared to input size.

All IRS warehouses are found in the vicinity of the origin, while all larger facilities exhibit DRS. One warehouse, with a relatively large FTE headcount and moderate floor space as well as few SKUs is an exception. In the interest of distinguishability, the axes were cut off at 150 FTEs, 50,000 $m^2$ and 80,000 SKUs, which excludes 15 warehouses, 7 of which with DRS and 4 with IRS. Out of the 4 IRS warehouses, 3 employ a very similar input and output mix, with high FTE counts, but relatively low other inputs, all with an efficiency score of less than 0.5. Aside from this particular mix of inputs and outputs, all other combinations are subject to DRS, which confirms the intuition by scholars in the field.

### 5.1.3 Scale efficiency

Scale inefficiency measures the inefficiency not attributable to inefficient input usage, but to sub-optimal plant size. Scale efficiency of a warehouse is defined as its DEA-CRS efficiency score – calculated by solving problem (1) – divided by its DEA-VRS efficiency score – calculated by solving the the same problem but including as additional constraint $\sum_{k'=1}^{K} \lambda_{k'} = 1$. As under VRS each facility is evaluated against its most favorable scale, VRS-based inefficiency can be attributed to pure operational aspects. The DEA-VRS score is also said to measure pure (input orientated) technical efficiency.

Under VRS, 46 warehouses are considered efficient in 2017 and 51 in 2012; with average efficiency scores of 0.78 in 2017 (standard deviation of 0.24) and 0.82 in 2012 (0.23). As under VRS more warehouses are efficient, there are more dominant peers. Because of this, the absolute number of times an observation acts as a peer in the

VRS case, compared to CRS, is reduced. Nevertheless there is consistency in the results since almost the same set of warehouses are found to be most often selected as peers in both years; that is, in 2017 WH#73 is peer for 35 inefficient counterparts, WH#34 for 35, and WH#33 for 25, while again others can be considered as extreme observations since they do not act as a reference for any inefficient warehouse (e.g., WH#26 or WH#66).

Table 6 reports that the average scale efficiency in 2017 was 0.88 with a standard deviation of 0.17 (0.92 and 0.13, respectively, in 2012). Scale inefficiency appears to be a considerable factor in the overall inefficiency of warehouses. The 0.68 technical efficiency in 2017 is caused by inefficient operations (0.78) and scale inefficiency (0.88). In 2012, both scale efficiency and operational efficiency were slightly higher (0.92 and 0.82 respectively). Based on this sample, the inefficiency split between operational and scale inefficiency is approximately 2:1 ((1-0,78):(1-0,88)), a ratio that has not been found previously in the literature. In other words, a third of the warehouses' inefficiency in this sample stems from operating at suboptimal scales.

Table 6: DEA efficiency scores.

| | 2012 | | | 2017 | | |
|---|---|---|---|---|---|---|
| | ITE | Pure TE | Scale efficiency | ITE | Pure TE | Scale efficiency |
| Minimum | 0.23 | 0.24 | 0.42 | 0.23 | 0.25 | 0.38 |
| Average | 0.75 | 0.82 | 0.92 | 0.68 | 0.78 | 0.88 |
| Maximum | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| Standard dev. | 0.24 | 0.23 | 0.13 | 0.25 | 0.24 | 0.17 |

## 5.2 Cross-efficiency results by methods

### 5.2.1 Ranking distributions, normality and similarity

We now present the results of applying the cross-efficiency methods discussed in section 3. We implemented the *classic* Sexton *et al.* (1986) approach as formulated in expression (8), the Sexton-*linear* surrogate approach as in expression (9), the *multiplicative* approach proposed by Cook and Zhu (2014) as formulated in expression (13), and the *game*-theoretic approach developed by Liang *et al.* (2008), here represented by expression (17).[24] All the models were run in their benevolent setting, 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.

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 approach are significantly lower (av-

[24]We also implemented the Sexton-*ratio* approach as defined by expression (10), and obtained minor differences with the Sexton-linear approach of expression (9). Although these results are contained in our tables and figures an explicit discussion in the text was deemed unnecessary. 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 five methods in both years are available upon request and will be accessible in online appendices.

erage = 0.360 and 0.362 for the first two models and 0.026 for the multiplicative approach, with standard deviations 0.177, 0.181 and 0.111, respectively).The game theoretic approach resulted in the highest average score of 0.535 (standard deviation 0.219) and finished after 32 iterations.[25] Table 7 reports the descriptive statistics of the cross-efficiency scores.

Table 7: Cross-efficiency scores, 2017.

| | Classic | Ratio | Linear | Multiplicative | Game |
|---|---|---|---|---|---|
| Minimum | 0.090 | 0.087 | 0.089 | 0.000 | 0.164 |
| Average | 0.360 | 0.362 | 0.363 | 0.026 | 0.535 |
| Maximum | 0.894 | 0.890 | 0.885 | 0.911 | 1.000 |
| Standard dev. | 0.177 | 0.181 | 0.177 | 0.111 | 0.219 |

Using the one-sample Kolmogorov-Smirnov (KS) test, the normal distribution assumption could be rejected at $p < 0.01$. The score distributions can be visually inspected in Figure 5. Most scores of the multiplicative approach are below 0.01, which most likely is due to the exponential nature of the calculation, heavily emphasizing efficient observations over inefficient ones, which leads to stark score differences compared to additive DEA calculations.

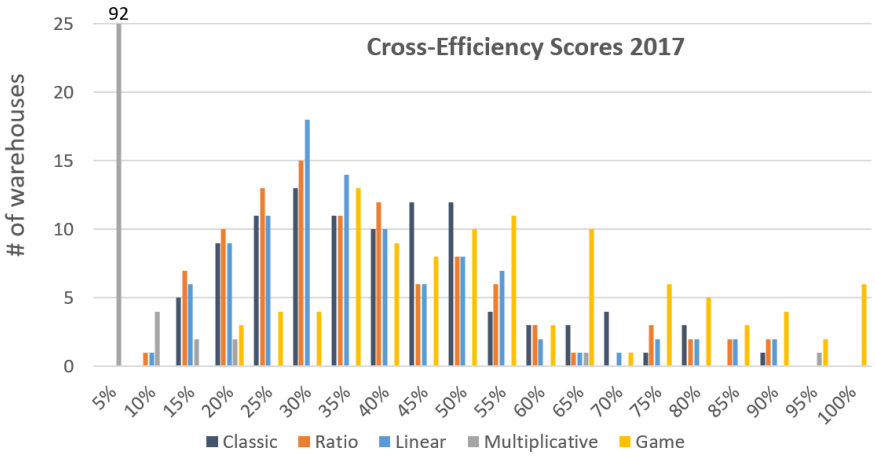


Figure 5: Cross-efficiency score distribution 2017 per method.

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

[25]In the intermediate runs two warehouses were obtained with cross-efficiency scores of 1. This is surprising, as a cross-efficiency method is designed to break the tie among efficient firms. However, as discussed in the literature review, this method does not follow the classic cross-efficiency path. Rather, it iterates through "rounds" in which the observations find Nash-equilibria for their weights and efficiency scores until those converge over time. Because in the preliminary set, the efficient warehouses employed mutually exclusive inputs in their optimal solution (and consequently had zero weights on all inputs the other facility uses), it is conceivable to find several "efficient" observations.

in 2017, except for the more similar Sexton-classic and -ratio methods in 2012. It is then clear that the choice of method is not neutral, leading to different rankings and presentations of warehouse performance within the industry.

The most critical question is of course whether the various methods result in comparable rankings. After all, managers are less interested in their particular cross-efficiency scores, but rather how their facilities behave comparatively and who is best-in-class. To give an impression of the different rankings, Table 8 reports the ranking of the Top 10 warehouses by cross-efficiency score for the five methods. Warehouses in bold signify that they are found in the Top 10 by all five methods.

Table 8: Top 10 warehouses, by cross-efficiency score and method, 2017.

| Rank | Classic | Ratio | Linear | Multiplicative | Game |
|---|---|---|---|---|---|
| 1 | **#098** | **#098** | **#098** | **#067** | #049 |
| 2 | **#050** | **#050** | **#050** | **#050** | **#098** |
| 3 | **#028** | **#028** | #049 | **#028** | **#050** |
| 4 | #049 | #049 | **#028** | **#027** | **#028** |
| 5 | **#104** | **#104** | **#104** | **#098** | **#104** |
| 6 | **#066** | **#066** | **#066** | **#066** | **#066** |
| 7 | **#027** | **#027** | **#067** | **#104** | **#027** |
| 8 | **#067** | **#067** | **#027** | #006 | **#067** |
| 9 | #041 | #041 | #041 | #115 | #107 |
| 10 | #107 | #040 | #040 | #040 | #041 |

It is noteworthy that already the Top 3 are not identical between the methods and that only 7 warehouses in 2017 and 5 in 2012 were ranked in the Top 10 by all methods. Still, the three Sexton-based methods exhibit very similar rankings, with warehouse ranks differing by usually 1-2 places. The game-theoretic approach ranks the Top 10 similar to the Sexton approaches (9 out of 10 facilities constituting the Top 10 are the same), whereas the ranking by the multiplicative approach differs considerably already in the first few warehouses.

Kendall's $\tau$ correlations were calculated for all method pairs. The three Sexton approaches show correlations of over 0.94 in both years, indicating almost identical rankings. Similarly, the Sexton-based methods and the game-theoretic approach correlate by more than 0.82 in both years. The correlations between the multiplicative approach and the other four approaches are considerably lower, namely in the range (0.63 - 0.68). The correlation between the cross-efficiency rankings and the first-stage $ITE$ ranking is also presented. As the latter yields 26 efficient warehouses, an additional ranking was compiled, based on super-efficiency (SE) scores following Andersen and Petersen (1993). The correlations with these two rankings follow the same pattern. The lowest corresponds to the multiplicative model (0.53), followed by the relatively low Sexton methods (0.64), and then by the moderately large game-theoretic approach.

Table 9: Kendall's $\tau$ for cross-efficiency rankings, 2017.

| | Classic | Ratio | Linear | Multiplicative | Game | $ITE$ | SE |
|---|---|---|---|---|---|---|---|
| Classic | 1 | | | | | | |
| Ratio | 0.96 | 1 | | | | | |
| Linear | 0.98 | 0.95 | 1 | | | | |
| Multiplicative | 0.67 | 0.68 | 0.67 | 1 | | | |
| Game | 0.82 | 0.82 | 0.83 | 0.63 | 1 | | |
| $ITE$ | 0.64 | 0.64 | 0.65 | 0.53 | 0.8 | 1 | |
| Super Efficiency (SE) | 0.64 | 0.64 | 0.65 | 0.53 | 0.8 | 0.95 | 1 |

Note: All $p$-values are $< 0.01$.

Based on these results in the next two subsections the Sexton-classic method will be employed.

## 5.3 Cross-efficiency results by warehouse clusters

This section reports our analysis of the relationship between cross-efficiency scores and warehouse characteristics, other than inputs and outputs. To start with, it is tested whether there are differences between the three warehouse clusters previously identified: Construction plus engineering, Consumer goods, and Food and groceries. To achieve this, a Kruskal-Wallis (KW) analysis was performed to test whether pairwise the cross-efficiency rankings between the three sectors are different.[26] In the overall sample the warehouses were ranked from 1, with the highest cross-efficiency score, to 102, with the lowest score. The test rejected the hypothesis that the ranks in the three clusters originate from the same distribution, with $p$ values 0.04 in 2017 and 0.03 in 2012.

To understand the differences between the groups better, Wilcoxon signed rank (WSR) tests were performed for each cluster against the remainder of the warehouses in the sample. Based on this analysis, it became evident that solely the Construction plus engineering cluster is statistically different (less efficient) from the rest. In both years, it exhibits a lower average rank – 65 in 2017 – than the other clusters and the overall sample. Although the other two clusters exhibit slightly higher average ranks than the overall sample, these differences are not significant. The detailed results are contained in Table 10.

[26]The KW test is a nonparametric test that does not assume a normal distribution of the residuals, and can be used to determine if there are statistically significant differences between two or more groups of equal or unequal size.

Table 10: Cluster characteristics, 2017.

| | Overall sample | Construction plus engineering | Consumer goods | Food and groceries |
|---|---|---|---|---|
| Observations | 102 | 20 | 19 | 19 |
| Minimum rank | 1 | 6 | 2 | 4 |
| Maximum rank | 102 | 96 | 98 | 101 |
| Average rank | 51.50 | 65.00 | 47.74 | 45.00 |
| Std. dev. of ranks | 29.44 | 28.19 | 26.44 | 26.36 |
| WSR $p$ value | | 0.02 | 0.54 | 0.29 |

Detailed analyses of the individual warehouses suggest that the worse than average cross-efficiency in the Construction plus engineering cluster is partially attributable to the high prevalence of spare-part facilities in this cluster that stock items for long periods and have lower output levels compared to their size. Five of the 20 Construction and engineering warehouses are focused on spare-part operations, while neither Consumer goods nor Food and groceries contain spare-part warehouses.

Performing the same analyses to test whether the position of a warehouse in the value chain has an impact on its cross-efficiency score we found that in 2017 there is no statistical difference between Production, Wholesale, and Retail warehouses. For 2012, the higher than average score for Retail is significant at the 5.8% threshold. However, given the low sample size of only 14 retail warehouses this finding should be verified with a larger data set.

Table 11: Value chain position characteristics, 2017.

| | Overall sample | Production | Wholesale | Retail |
|---|---|---|---|---|
| Observations | 102 | 27 | 61 | 14 |
| Minimum score | 0.09 | 0.09 | 0.12 | 0.18 |
| Maximum score | 0.89 | 0.89 | 0.87 | 0.82 |
| Average score | 0.36 | 0.39 | 0.34 | 0.38 |
| Std. dev. of scores | 0.18 | 0.22 | 0.16 | 0.17 |
| KW $p$-value | | | 0.6704 | |

Finally, the impact of ownership type on cross-efficiency scores was also analyzed through KW tests. It was found that neither in 2017 nor in 2012 there is a statistical difference between the In-House, 3PL-Dedicated, and 3PL-Multiple facilities. In both years though, the In-house warehouses observed the highest efficiencies, while 3PL-operated warehouses were less efficient. At the same time, the scores of both 3PL groups exhibited lower standard deviations. Both findings can be explained by noticing that the less efficient facilities may be those with a lower degree of operational specialization, as specifically tailored (In-house) warehouses do not need to exhibit the operational multi-tasking flexibility associated with third-party logistics, which is subject to ever changing contract conditions and uncertain temporal

continuity.

Table 12: Ownership characteristics, 2017.

| | Overall sample | In-House | 3PL-Dedicated | 3PL-Multiple |
|---|---|---|---|---|
| Observations | 102 | 58 | 10 | 34 |
| Minimum score | 0.09 | 0.09 | 0.21 | 0.12 |
| Maximum score | 0.89 | 0.89 | 0.55 | 0.58 |
| Average score | 0.36 | 0.39 | 0.33 | 0.33 |
| Std. dev. of scores | 0.18 | 0.22 | 0.10 | 0.11 |
| KW $p$-value | | | 0.8949 | |

We conclude that the cross-efficiency scores are not significantly different between activity clusters, value chain positions, or ownership types. It is reassuring that, except for a few exceptions, managers are not handicapped by characteristics which are not at their own discretion.

## 5.4 Research hypotheses findings

The hypotheses introduced in Section 4.2 are evaluated on the overall sample of 102 warehouses. Table 13 shows the main results. The second column contains standardized coefficients of an ordinary least squares (OLS) regression performed on the cross-section of warehouses in 2012. The dependent variable is the Sexton-classic cross-efficiency score and the explanatory variables are the input and output variables. The third column concerns the situation in 2017. The fourth column merits special attention. The explanatory variables are now the rates of change, between 2012 and 2017, of the input and output variables. The cross-efficiency scores of the two years cannot be cardinally compared panelwise, since they are defined against the backdrop of two different technological frontiers. Therefore as dependent variable the change of a warehouses' ranking position was taken.[27]

The standardized coefficients appear to be statistically significant except that for Error free percentage. The intercept is 0 and therefore not displayed in the table.

[27] As for the OLS regression strategy, we remark that cross-efficiency scores are not censored except in the game-theoretic approach (Table 7). Despite this, Hoff (2007) found that OLS predicted the ranks slightly better than three commonly used alternative models (Tobit regression, Papke-Wooldridge approach, unit-inflated beta model). We estimated alternative regressions controlling for the percentage of floor space dedicated to cold storage, cluster category, value chain position, and ownership type, the last three included as dummy variables. However, in accordance with the results reported in the previous section, all these variables appeared to be uncorrelated with the cross-efficiency scores. These simple results based on OLS indicate that it is not worth pursuing a two-stage DEA analysis in which bootstrapping techniques are invoked to estimate robust cross-efficiency scores, which then would be regressed on the above explanatory variables. Following Simar and Wilson (2007), had these variables revealed themselves as significant drivers of cross-efficiency, then bootstrapping procedures for a two-stage implementation, involving the initial cross-efficiency estimation based on data resampling and bias correction, and subsequent econometric analysis using maximum-likelihood truncated regressions, would have been warranted. For further details we refer the reader to algorithms #1 and #2 in their paper.

Floor space exhibits the lowest (negative) correlation with the cross-efficiency score, followed by SKUs. The explanatory fit of the model (measured by adjusted $R^2$, $F$-value, as well as additional goodness-of-fit statistics) is satisfactory. Thus we rely on these results to discuss managerial questions and research hypotheses.

Table 13: Standardized regression coefficients.

| | Sexton-classic score 2012 | Sexton-classic score 2017 | Change of ranking position |
|---|---|---|---|
| Floor space | -0.31*** | -0.16* | 0.05 |
| FTEs | -0.49*** | -0.54*** | -0.19** |
| SKUs | -0.35*** | -0.27*** | -0.57*** |
| Automation | -0.43*** | -0.51*** | -0.67*** |
| Order lines | 0.69*** | 0.66*** | 0.61*** |
| Error free percentage | 0.05 | 0.06 | 0.15* |
| Order flexibility | 0.23*** | 0.09 | 0.24*** |
| Special processes | 0.26*** | 0.23*** | 0.27*** |
| Adjusted $R^2$ | 0.54 | 0.49 | 0.52 |
| RMSE | 0.68 | 0.72 | 0.69 |
| F-value | 15.70 | 12.90 | 14.70 |

Note: $p$-values: $^*p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

The hypotheses 1a and 1b that Automation, both in levels and rates of change, positively correlates with efficiency should be rejected. In the two years considered the correlation appears to be significantly negative. Likewise, the rank correlation between the two variables is -0.61 in 2012 and -0.66 in 2017. The same negative relation holds between the change of automation between 2017 and 2012 and the change in ranking position, with a regression coefficient of -0.67. Note that given the ratio (dual) definition of $ITE(x^k, y^k)$ as presented in expression (2), inputs (in the denominator) and outputs (in the numerator) are expected to correlate negatively and positively, respectively, with the cross-efficiency scores. Therefore, confirmation of hypotheses 1a and 1b would have required reversals in the expected signs. These, however, did not materialize.

Inspecting the relationship between automation and size, it is observed that higher automation scores are observed in the larger warehouses. This finding supports the hypothesis that larger warehouses use more automation, because otherwise they could not stay competitive. Simultaneously, they are less technically efficient, as they are more likely subject to decreasing returns to scale as argued in Section 5.1. This interpretation would explain why automation is evidently increasingly used, although it does not seem to positively correlate with efficiency. Above a certain size, however, automation conditionally does play a positive role.

Despite the overall negative correlation between automation and cross-efficiency, there are certain differences among the three warehouse clusters. While the average automation score of Consumer goods warehouses is slightly below the average (6.58 vs. 6.83), Food and groceries warehouses have a mean automation score of 8.21. Performing the same regressions as presented in Table 13 for each of the three clus-

ters, Consumer goods warehouses exhibits the highest negative regression coefficient (-0.89 at a $p$-level of 0.01), the highest change of automation level, as well as the highest *number* of warehouses that changed their automation level.[28]

Contrary to the automation case, the hypotheses 2 and 3, stating that SKUs and FTEs correlate negatively with efficiency is supported by the sample. The same holds for the temporal rates of change. In 2017 the overall sample exhibits a standardized regression coefficient of -0.27 for SKUs (Table 13). The same magnitude is obtained for the clusters, though only significant for Construction plus engineering. This is certainly in accordance with the previously identified existence of decreasing returns to scale.

For SKUs as well as FTEs, the clusters exhibit standard deviations greater than or equal to the means. On average, 59 FTEs work per facility, with Construction plus engineering facilities employing the most (78). The regression coefficient of the cross-efficiency score on FTEs is –0.54 for the overall sample (Table 13), while it is –2.69 for the Construction plus engineering cluster. Ninety percent of the warehouses in the entire sample changed the number of FTEs from 2012 to 2017, on average by 32%. Food and groceries warehouses spiked with an average increase of 43%.

Finally, hypothesis 4a that Floor space and efficiency correlate positively is also rejected by the results in Table 13. Thus the hypothesis that larger warehouses are more efficient due to increasing returns to scale is also rejected, as previously discussed in section 5.1. This result extends over time, as hypothesis 4b, suggesting a positive correlation between floor space change and efficiency change, is not supported by our data. The two variables have a correlation coefficient close to zero and not significant. Similar results hold or the three clusters distinguished.

# 6 Evaluation of cross-efficiency methods

## 6.1 Method comparison dimensions

In this section we focus on the relative merits of the four cross-efficiency methods employed, taking into consideration their differences in implementation, usage and results according to the dimensions presented in Table 14. Each method is given a score on every dimension, with a value between 1 and 5, with 5 signifying the

[28] Further field research was carried out to investigate this conjecture. The authors visited the largest food warehouse in the sample – the Food and groceries cluster exhibits the highest average automation score – and discussed the reasons for automation investment decisions with the management team. They informed us that the high level of automation installed in their facility was necessary, because the tens of thousands perishable units they receive and ship every day could not be handled by purely manual labor at competitive costs. But even more important, conveyor belts, automated stackers, labeling machines etc. ensure better quality and higher consistency. In such a sensitive environment, where slight mistakes render the product obsolete and hygiene regulations dictate process design, a fully automated technology is the best option. In this particular case, the high quality focus and the perishable assortment lead to a low cross-efficiency rank, yet all-manual operations would likely rank the warehouse worse. Because of this, the relation between automation and cross-efficiency is not causal, but influenced by operational necessities of specific industries and warehouse sizes.

best.[29] First and foremost, whether the sets of results, obtained through the various methods, are significantly different is determined by performing Wilcoxon signed rank tests. If indeed different then, based on the scores assigned at the various dimensions, recommendations are provided on when to use which method. As the Sexton-classic and its linear surrogate are not statistically different, the latter is dropped from the analysis.

For the calculation of the discriminatory power of the various methods, two separate approaches were followed: 1) The four methods were applied in their standard form as described in section 3. 2) The (second-stage) cross-efficiency scores were calculated employing only the weights of warehouses with a first-stage efficiency score of 1. The intuition is that when calculating cross-efficiency scores in the standard way, all the sets of warehouse weights are valued equally. However, inefficient warehouses exhibit weights that closely correspond to the weights of their efficient peers at the frontier hyperplane. Thus, including inefficient observations in the cross-efficiency analysis would amount to multiple counting the weights of the dominant warehouses.

## 6.2 Ranking of methods: Capacity and sensitivity

The evaluation of the various methods according to the dimensions outlined above is presented in Table 15. The explanation of the highest and lowest scores immediately follows.

Table 15: Ratings of cross-efficiency methods across dimensions.

| | Linear | Classic | Multiplicative | Game |
|---|---|---|---|---|
| Meth. proximity to plain DEA | 3 | 4 | 2 | 4 |
| Implementational ease | 5 | 4 | 4 | 2 |
| Extendability | 4 | 4 | 2 | 3 |
| Discriminatory properties | 5 | 5 | 2 | 4 |
| Sensitivity to changes of scale | 5 | 5 | 2 | 5 |
| Sensitivity to erroneous data | 5 | 5 | 3 | 5 |
| Sensitivity to peers deletion | 4 | 4 | 1 | 3 |
| Total | 31 | 31 | 16 | 28 |

*Methodological proximity to plain DEA*: Obviously, the Sexton-classic approach stays closer to plain DEA than the Sexton-linear approach. However, cross-efficiency results show a correlation with plain DEA results of just 0.6. On the contrary, the Game results, based on an iterative process and pair-wise optimization rather than global optimization, exhibit the highest correlation with plain DEA, namely 0.8. The Multiplicative approach is assigned the lowest score. The exponential transformation of the objective function results in the lowest correlation with plain DEA, namely approximately 0.5.

[29] This procedure is useful for comparing methods across single dimensions, but it is not intended for computing an aggregate score per method, as different situations may lead to different choices. Nevertheless, a simple aggregate score is calculated to ease the comparison of methods.

Table 14: Comparison dimensions for the cross-efficiency methods.

| Dimension | Description |
|---|---|
| Methodological proximity to plain DEA | Tests how close a cross-efficiency method follows the logic of the 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 data set. 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 described in the literature, which allows tweaking the basic method to the requirements of an individual project. |
| 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 peer 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. |

*Implementational ease*: The Sexton-linear approach is a simple linear program that can be directly implemented in any spreadsheet with mathematical programming tools. It results in almost instant solutions. Although the Sexton-classic approach is nonlinear, and the multiplicative approach involves a number of constraints growing quadratically with the number of observations – more than thousand for the 102 warehouses –, both of them solve in a few seconds with a standard quad-core 3.8 GHz processor. By far, the most demanding and slowest method is the game-theoretic, as the pairwise comparisons require calling the optimizer thousands of times. (In our case it took over 70 minutes to solve the program.)

*Extendability*: Throughout the literature of DEA applications, the Sexton-linear and -classic approaches are the most popular and easiest to adopt in empirical studies. The multiplicative approach is relatively new, and although possible extensions were mentioned by Cook and Zhu (2014) only a limited number of citations could be found. For the game-theoretic approach there appears to be a considerable number of publications, however mainly co-authored by its pioneers. In general this approach does not find application among DEA pundits.

*Discriminatory properties*: A cross-efficiency method is intended to break the ties between the warehouses found efficient at the first DEA stage (26 and 35 in 2017 and 2012, respectively). To get an idea about the performance of the methods Table 16 compares the (mean of 2012 and 2017) range, standard deviation, and kurtosis of the cross-efficiency score distributions for these efficient warehouses. The upper panel uses thereby the data of all the 102 warehouses, the lower panel uses only the data of the efficient warehouses.

Table 16: Characteristics of cross-efficiency score distributions for warehouses found efficient at the first stage, 2012 and 2017.

| Based on all warehouses | Linear | Classic | Multiplicative | Game |
|---|---|---|---|---|
| Mean range | 0.64 | 0.65 | 0.92 | 0.53 |
| Mean standard deviation | 0.17 | 0.17 | 0.20 | 0.14 |
| Mean kurtosis | 0.20 | 0.05 | 2.28 | -0.16 |

| Based on efficient warehouses | Linear | Classic | Multiplicative | Game |
|---|---|---|---|---|
| Mean range | 0.63 | 0.65 | 0.88 | 0.36 |
| Mean standard deviation | 0.17 | 0.18 | 0.20 | 0.10 |
| Mean kurtosis | -0.34 | -0.43 | 3.01 | 0.14 |

Comparing the sets of numbers in the table we firstly see that with respect to range and standard deviation of the final cross-efficiency scores there is not much to choose between the four methods. The difference is in the kurtosis. The extremely high magnitude for the multiplicative method means that the final scores are strongly clustered, which makes discrimination between the warehouses difficult. Secondly, it turns out that basing the cross-efficiency calculation only on the data of the first-stage efficient warehouses does hardly change range and standard deviation of the final scores. It is here again the kurtosis that matters: for both Sexton-

based methods the kurtosis changes from positive to negative, for the game-theoretic approach from negative to positive, but for the multiplicative approach the kurtosis becomes higher.

A disadvantage of the game-theoretic approach appears to be its propensity to rate up to two warehouses as equally cross-efficient; e.g., with unitary efficiency scores. This happens when the warehouses are both first-stage efficient and employ mutually exclusive inputs in their optimum. In this case the advantage of the cross-efficiency method of providing a unique ranking without ties, based on different individual scores, is lost.

*Sensitivity to scale changes*: The outcome of any DEA-based method is generally sensitive to which decision making units determine the production frontier. For the scale sensitivity test the inputs of a random selection of 10% of the warehouses were scaled by factors 0.1, 0.2, 0.5, 2, 5, or 10, respectively. This serves to reflect industries with volatile inputs, or models expressing inputs in monetary units. For each of the factors 100 runs were performed per model, except for the game-theoretical approach, where to save on computing time only 10 runs were performed. The resulting cross-efficiency scores from these simulations were then compared with the original scores.

The two Sexton-based methods and the game-theoretic method exhibited a low sensitivity to scale changes – the difference between simulated and original scores being always lower than 20%. For the multiplicative method the differences were three to four times higher.

*Sensitivity to erroneous entries*: In this case all inputs were scaled by a random value in the 75% - 125% range, for 5%, 10%, 25%, 50%, 75% and 100% of the warehouses. This serves to reflect situations like the present study where data is collected through questionnaires and therefore subject to various kinds of human error. Depending on consistency checks during the collection processes, the required granularity, and the type of data requested, the acquired data may vary in exactness. This test checks how the cross-efficiency methods handle such imprecisions.

For each simulation the relative difference between obtained and original score was calculated and averages over the simulation runs compared. It appeared that both Sexton-based methods and the game-theoretic approach were more stable than the multiplicative approach, with percentages of 10 and 30 respectively.

*Sensitivity to peers deletion*: For this test, 1-6 first-stage efficient warehouses were randomly dropped from the data set. This would simulate what happens when, for instance, warehouses are reclassified, or when facilities are excluded because they are deemed non-representative or "maverick". Eliminating those peers is expected to have some impact on the resulting cross-efficiency scores of the non-peers.

Once again the two Sexton-based methods and the game-theoretic approach exhibited relatively small, whereas the multiplicative approach exhibited a high sensitivity.

Looking at the last row of Table 15 it is clear that the two Sexton-based methods are performing best. A discussion of what method is recommendable in which situation follows in the conclusions, where the specific characteristics of our warehouse study will be taken into account.

# 7 Summary and conclusions

In this study we applied four representative cross-efficiency methods to measure warehouse performance, using data collected in 2017. Based on 102 warehouses operating in the Netherlands and Belgium, and a set of relevant inputs and outputs, it was found that the industry is subject to decreasing returns to scale. From a managerial perspective, the size recommendations in section 5.1 are of relevance, as it was shown that 1/3 of inefficiencies in warehousing stems from scale inefficiency and 2/3 from operational inefficiency. Also, in line with previous research on the warehouse sector, all the four input factors correlate significantly and negatively with cross-efficiency: Automation with -0.66, Floor space with -0.43, SKUs (assortment size) with -0.78, and FTEs (number of employees) with -0.37; these data relate to 2017 but the same orders of magnitude and signs hold in 2012.

The first finding contradicts the intuition that automation and cross-efficiency would correlate positively. One possible explanation for this is that the automation score did not include common technologies found in warehouses, such as reach trucks, racking systems, etcetera. Hence, warehouses relying heavily on classic warehouse setups, with limited or no automation technology asked for in the survey (like AGVs, automated stackers, RFID technology, or conveyor belts) can be very efficient, yet get a low automation score. At the same time, when looking at the automation scores at most efficient scale sizes, low to moderate automation levels are prevalent.

Across the entire set of warehouses it was found that the remaining input factors, floor space, SKUs, and FTEs, correlate negatively with the cross-efficiency score. These results did not change substantially when running the cross-efficiency models for the warehouse clusters separately. In conjunction with the input and output mix comparison of the industries, the key learning for practitioners is to consider size as well as the choice of input and output, without solely focusing on the overall cross-efficiency scores.

When contrasting the four cross-efficiency methods, it appeared that the Sexton-classic and -linear methods, followed by the game-theoretic method, beat the multiplicative method across almost all dimensions.

Choosing between the two Sexton-based methods depends a bit on the preference of the user and the context of the application. While the linear method is slightly quicker solvable, the classic framework as implemented in this paper approximates the first-stage DEA program methodologically more closely. From the computational point of view the game-theoretic method is distinctively disadvantageous in terms of required computer time.

All in all, we recommend Sexton-classic as the default model to use.

We conclude with pointing out some limitations of this study and indicating areas for further research.

First, while a dataset of over 100 warehouses fulfills all DEA minimum requirements, a larger dataset would be preferable. For the study of industrial or bespoke subsets a larger sample is necessary. Next to that, the temporal analysis was limited by the use of recollection from respondents about a warehouse's situation in the recent past. A true panel survey, preferably executed over a longer time span than five

years, could help in more accurately grasping trends in warehousing performance.

Second, a combination of existing questions, experience, and personal judgment was used to develop the automation section of the questionnaire. A holistic framework for capturing mutually exclusive, collectively exhaustive automation components is not available, which limits the precision of the automation scoring.

Third, the current research on cross-efficiency measurement with variable returns to scale (VRS) is at too early a stage to reliably use the resulting scores. As the warehouses in our sample appear to operate both in regions of increasing and decreasing returns to scale, our cross-efficiency scores can probably not sufficiently account for scale (in)efficiency as they are derived under the assumption of CRS.

Further research is needed on the drivers of warehouse efficiency, such as specific product categories, value chain positions, and relevant inputs and outputs. As the industrial development is fast, a biennial efficiency survey would be of interest for researchers and professionals. Alternatively, a web-hosted solution like the one proposed by Johnson and McGinnis (2011) could be programmed to provide warehouse managers with an interface to submit operational data and have their warehouse ranked through a cross-efficiency computer code. This way, warehouse data could continuously be gathered, while participants receive results instantly.

From the methodological perspective a future focus on VRS-based cross-efficiency measurement along the lines indicated by Lim and Zhu (2015) would be highly welcome. A further topic is the introduction of the temporal dimension in cross-efficiency measurement so that productivity indices can be defined and it becomes possible to identify the role of efficiency change and technological change. One could also think of frameworks which allow more realistic representations of warehouse processes, such as network DEA, in which automation could be treated as an intermediate rather than an input or output.

# References


Aczél, J. and F. S. Roberts, 1989, "On the possible merging functions", Mathematical Social Sciences 17, 205-243.

Allen, R., Athanassopoulos, A., R. D. Dyson and E. Thanassoulis, 1997, "Weights restrictions and value judgements in Data Envelopment Analysis: evolution, development and future directions", Annals of Operations Research 73,13-34.

Andersen, P. and N. C. Petersen, 1993, "A procedure for ranking efficient units in data envelopment analysis", Management Science 39, 1261-1264.

Anderson, T. R., K. Hollingsworth and L. Inman, 2002, "The fixed-weighting nature of a cross-evaluation model", Journal of Productivity Analysis 17, 249-255.

Avkiran, N. K., 2006, "Productivity analysis in the service sector with Data Envelopment Analysis". Available at SSRN: ssrn.com/abstract=2627576.

Balk, B. M., 2008, Price and Quantity Index Numbers: Models for Measuring Aggregate Change and Difference (Cambridge University Press, New York).

Banker, R. D., H. Chang and W. W. Cooper, 1996, "Equivalence and implementation of alternative methods for determining returns to scale in data envelopment analysis", European Journal of Operational Research 89, 473-481.

Banker, R. D., A. Charnes and W. W. Cooper, 1984, "Some models for estimating technical and scale inefficiencies in Data Envelopment Analysis, Management Science 30, 1078-1092.

Banker, R. D. and R. M. Thrall, 1992, "Estimation of returns to scale using data envelopment analysis", European Journal of Operational Research 62, 74-84.

Charnes, A., W. W. Cooper and E. Rhodes, 1978, "Measuring efficiency of decision making units", European Journal of Operational Research 2, 429-444.

Charnes, A., W. W. Cooper, L. Seiford and J. Stutz, 1982, "A multiplicative model for efficiency analysis", Socio-Economic Planning Sciences 16, 223-224.

Charnes, A., W. W. Cooper, L. Seiford and J. Stutz, 1983, "Invariant multiplicative efficiency and piecewise Cobb-Douglas envelopments", Operations Research Letters 2, 101-103.

Chen, J.-X., 2012, "A comment on DEA efficiency assessment using ideal and anti-ideal decision making units", Applied Mathematics and Computation 219, 583-591.

Cook, W. D. and J. Zhu, 2014, "DEA Cobb-Douglas frontier and cross efficiency", Journal of the Operational Research Society 65, 265-268.

Cook, W. D. and J. Zhu, 2015, "DEA cross-efficiency", in Data Envelopment Analysis, edited by J. Zhu, International Series in Operations Research & Management Science 221 (Springer Science+Business Media, New York).

Cooper, W. W., J. L. Ruiz and I. Sirvent, 2011, "Choices and uses of DEA weights", in Handbook On Data Envelopment Analysis, edited by W. W. Cooper, L. M. Seiford and J. Zhu (Springer Science+Business Media, New York).

De Koster, M. B. M. and B. M. Balk, 2008, "Benchmarking and monitoring international warehouse operations in Europe", Production and Operations Management 17, 175-183.

De Koster, M.B.M. and P.M.J. Warffemius, 2005, "American, Asian and third-party international warehouse operations in Europe: A performance comparison", International Journal of Operations & Production Management 25(8), 762-780.

De Koster, M. B. M., T. Le-Duc and N. Zaerpour, 2012, "Determining the number of zones in a pick-and-sort order picking system", International Journal of Production Research 50, 757-771.

De Vries, J., M. B. M. de Koster and D. Stam, 2016, "Exploring the role of picker personality in predicting picking performance with pick by voice, pick to light and RF-terminal picking", International Journal of Production Research 54, 2266-2274.

Doyle, J. R. and R. H. Green, 1994, "Efficiency and cross-efficiency in DEA: Derivations, meanings, and uses", Journal of the Operational Research Society 45, 567-578.

Doyle, J. R. and R. H. Green, 1995, "Cross-evaluation in DEA: Improving discrimination among DMUs", Infor 33, 205-222.

Dyson, R. G., R. Allen, A. S. Camanho, V. V. Podinovski, C. S. Sarrico and E. A. Shale, 2001, "Pitfalls and protocols in DEA", European Journal of Operational Research 132, 245-259.

Emmett, S., 2005, Excellence in Warehouse Management (John Wiley & Sons, West Sussex).

Eurostat, 2017, "Annual detailed enterprise statistics for services (NACE Rev. 2 H-N and S95)", Luxemburg, appsso.eurostat.ec.europa.eu/nui/show.do.

Faber, N., M. B. M. de Koster and A. Smidts, 2013, "Organizing warehouse management", International Journal of Operations & Production Management 33, 1230-1256.

Golany, B. and Y. Roll, 1989, "An application procedure for DEA", Omega 17, 237-250.

Gu, J, M. Goetschalckx, and L. F. McGinnis, 2007, "Research on warehouse operation: A comprehensive review", European Journal of Operational Research 177, 1-21.

Hackman, S. T., E. H. Frazelle, P. M. Griffin, S. O. Griffin and D. A. Vlasta, 2001, "Benchmarking warehousing and distribution operations: An input-output approach", Journal of Productivity Analysis 16, 79-100.

Hamberg, R. and J. Verriet, 2012, Automation in Warehouse Development (Springer, New York).

Hoff, A., 2007, "Second stage DEA: Comparison of approaches for modelling the DEA score", European Journal of Operational Research 181, 425-435.

Johnson, A. and L. F. McGinnis, 2011, "Performance measurement in the warehousing industry", IIE Transactions 43, 220-230.

Liang, L., J. Wu, W. D. Cook and J. Zhu, 2008, "The DEA game cross efficiency model and its Nash equilibrium", Operations Research 56, 1278-1288.

Lim, S. and J. Zhu, 2015, "DEA cross-efficiency under variable returns to scale", in Data Envelopment Analysis, edited by J. Zhu, International Series in Operations Research & Management Science 221 (Springer Science+Business Media, New York).

Lu, W.-M., and S.-E. Lo, 2007, "A benchmark-learning roadmap for regional sustainable development in China", Journal of the Operational Research Society 58, 841-849.

Muilerman, G.-J., 2001, Time-based logistics, PhD thesis, Delft University of Technology.

Pastor, J. T., J. L. Ruiz and I. Sirvent, 2002, "A statistical test for nested radial DEA models", Operations Research 50, 728735.

Ramón, N., J. L. Ruiz and I. Sirvent, 2010, "On the choice of weights profiles in cross-efficiency evaluations", European Journal of Operations Research 207, 1564-1572.

Saaty T. L., 1986, The analytic hierarchy process, (McGraw-Hill International Book Company, New York).

Sauermann, H. and M. Roach, 2013, "Increasing web survey response rates in innovation research: An experimental study of static and dynamic contact design features", Research Policy 42, 273-286.

Sexton, T. R., R. H. Silkman and A. J. Hogan, 1986, "Data Envelopment Analysis: Critique and Extensions", in Measuring Efficiency: An Assessment of Data Envelopment Analysis, edited by R. H. Silkman, New Directions for Program Evaluation 32 (Jossey-Bass, San Francisco/London).

Shen, W., Zhang, D., W. Liu and G. Yang, 2016, "Increasing discrimination of DEA evaluation by utilizing distances to anti-efficient frontiers", Computers & Operations Research 75, 163-173.

Simar, L. and P. W. Wilson, 2007, "Estimation and inference in two-stage, semi-parametric models of production processes", Journal of Econometrics, 136, 31-64.

Thanassoulis, E., M. C. S. Portela and R. Allen, 2004, "Incorporating value judgements in DEA", in Handbook on Data Envelopment Analysis, edited by W. W. Cooper, L. W. Seiford and J. Zhu (Kluwer Academic Publishers, Boston).

Thompson R. G., Singleton, F. D., R. M. Thrall and B. A. Smith, 1986, "Comparative site evaluations for locating a high-energy physics lab in Texas", Interfaces 16, 3549.

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.

Van der Gaast, J. P., 2015, Stochastic models for order picking systems, PhD Series in Research in Management #648, ERIM, Erasmus University Rotterdam, The Netherlands.

Wang, Y.M. and K.S. Chin, 2010, "Some alternative models for DEA cross-efficiency evaluation", International Journal of Production Economics 128, 332-338.

Wang, Q., R. McIntosh and M. Brain, 2010, "A new-generation automated warehousing capability", International Journal of Computer Integrated Manufacturing 23, 565-573.

Wu J., L. Liang and F. Yang, 2009a, "Determination of the weights for the ultimate cross-efficiency using Shapley value in cooperative game", Expert Systems with Applications 36, 872-876.

Wu J., L. Liang, F. Yang and H. Yan, 2009b, "Bargaining game model in the evaluation of decision making units", Expert Systems with Applications 36, 4357-4362.

Zhu, J., 2014, Quantitative Models for Performance Evaluation and Benchmarking (Springer International Publishing Switzerland).