www.palgrave-journals.com/jors/

Performance comparison of retailing stores using a Malmquist-type index

CB Vaz[1]* and AS Camanho[2]

[1]*Escola Superior de Tecnologia e Gestão, Instituto Politécnico de Bragança, Portugal; and* [2]*Faculdade de Engenharia, Universidade do Porto, Portugal*

This study develops a framework that combines different management science methods to provide insights concerning the performance of retailing stores. First, the framework enables to specify appropriate targets for stores of a retail network using data envelopment analysis. This involves comparing stores within homogenous groups, that is, supermarkets and hypermarkets. Second, the framework compares the overall performance of these two groups. This requires the combined use of a Malmquist-type index and statistical tests. This index is decomposed into sub-indices for comparing the differences between groups in terms of the efficiency spread in each group of stores and the productivity differences between the best-practice frontiers spanned by the benchmark stores from each group. The hypothesis tests are used to verify if the differences between groups captured by the sub-indices are statistically significant.

Journal of the Operational Research Society (2012) **63,** 631–645. doi:10.1057/jors.2011.63
Published online 13 July 2011

Keywords: data envelopment analysis; retail; efficiency; productivity; malmquist index; statistical tests

1. Introduction

This study aims to apply to a retail network, a framework that combines management science methods (data envelopment analysis (DEA), Malmquist indices and statistical tests) to provide insights concerning the performance of stores. First, a DEA model enables to specify appropriate targets for stores of a retailing network. This involves comparing stores of the same type, that is, supermarkets or hypermarkets, within their group. Second, a Malmquist-type index (MI) complemented by statistical hypothesis tests enables comparing the overall performance of those groups, which requires characterizing their productivity levels. The models and methods developed in this paper were motivated by the problems faced in a real-world organisation. Our aim was to design a new method to compare the performance of different types of stores, that can contribute to the enhancement of the operation of retailing organisations, guiding them towards continuous improvement.

From the company perspective, this research intended to contribute to the specification of appropriate sales targets for each store. At present, these targets are defined on a yearly basis by the planning and control department of the company. The sales potential of each store is determined on the basis of internal benchmarking to ensure that the targets defined are achievable. Assuring equity in this process is an issue that must be carefully considered. This has led the planning and control department of the company to classify the stores into homogenous groups that are analysed separately in the internal benchmarking process: the hypermarkets and the supermarkets. The criteria that differentiates the two groups is based on floor area and the size of the urban area where the stores are located. Internal benchmarking is carried out within each group to compare the stores performance. This involves the analysis of performance indicators, which include ratios of outcomes (eg, sales or profit) over resources (eg, number of employees or costs) and environmental conditions reflecting market potential (eg, population or competition). The main difficulty of this process is to identify fair benchmarks for a given store, such that the sales target defined is accepted by the store manager.

One of the limitations of using a set of indicators for performance appraisal is that they cannot be used in a straightforward manner to set targets. This is because each single indicator has to be compared with some benchmark value, without regarding the remaining aspects of the store activity that are not accounted for in that indicator. Although any particularly poor value of an indicator identifies an aspect of the store activity in special need of improvement, the target levels cannot be estimated with confidence, as achieving a target for one indicator may have implications on other dimensions of store activity. We believe that the use of enhanced productivity measurement methods, such as DEA

*Correspondence: CB Vaz, Escola Superior de Tecnologia e Gestão, Instituto Politecnico de Bragança, Campus Santa Apolonia, Apartado 134, Bragança 5301-857, Portugal.
E-mail: clvaz@ipb.pt

for benchmarking purposes in the retailing sector can contribute in a positive way to overcome this limitation.

The second objective of this paper concerns the comparison of productivity between the groups of supermarket stores and the hypermarket stores. From a strategic perspective, it is important to understand which type of store is more productive, and to quantify the productivity difference. This analysis also intents to explore if there are reasons for supporting the current option of the company to separate the internal benchmarking analysis of supermarkets from the analysis of hypermarkets.

For the groups comparison, we combine the use of a MI with statistical tests. The calculation of the MI is based on distance functions, which can be estimated using DEA models. The MI is usually applied to the measurement of productivity change over time, and can be multiplicatively decomposed into an efficiency-change index and a technological-change index. This paper uses a modified version of the MI, whose fundamental characteristic is to focus on group comparisons in a static setting. This index can be decomposed into sub-indices for comparing the efficiency spread in each group, and the productivity differences between the best-practice frontiers of each group. The evaluation of the efficiency spread within the groups gives an indication to what extent the performance of the stores is homogeneous, that is, if all supermarkets and hypermarkets are equally close to the best-practice levels observed within their own group, or if in one group the stores are closer to the frontier than in the other. The value of the sub-index comparing the frontiers' productivity quantifies the magnitude of the differences in the location of the best-practice frontiers, which can be relevant information for strategic purposes. The hypothesis tests are used to verify if the differences between groups captured by the sub-indices are statistically significant. From a managerial perspective, if the sub-index comparing the frontier productivity reveals statistically significant differences, it should be interpreted as an evidence that the benchmarking analysis should be done separately for each group of stores.

This paper is organised as follows. The next section provides a brief review of the literature on retailing performance. Section 3 describes the performance assessment methods used in this paper (DEA and Malmquist indices). It also describes the hypothesis tests proposed for exploring if the differences between the two types of stores analysed are statistically significant, as captured by the MI and its components. At the end of this section, we provide a brief demonstration of how these methods can be combined to assess performance using illustrative examples generated by a Monte Carlo simulation. Section 4 describes the contextual setting of the retailing sector and provides a brief description of the company used as a case study, whose stores operate in Portugal. Section 5 reports the results of the assessment and discusses their managerial implications. The last section concludes.

2. Review of literature on retailing performance

Efficiency and productivity are important issues in the retailing sector, which affect the performance of stores. Those concepts are sometimes used interchangeably in the retailing literature (such as in Kamakura *et al*, 1996; Donthu and Yoo, 1998). To clarify the distinction between the efficiency and productivity concepts that underlie this study, consider the case of production units, hereafter called decision-making units (DMUs) that use one input to produce one output. Productivity is defined as the ratio of output produced over input used by the DMU. Efficiency compares this ratio in each DMU with the best ratio observed among all DMUs. Thus, productivity is an absolute concept whereas efficiency is a relative measure. A DMU is efficient if it achieves the highest ratio observed in all DMUs analysed, and the magnitude of inefficiency reflects the level of underachievement in relation to the maximum productivity level observed in the sample.

Traditionally, in a retailing setting, ratio analysis, such as in Lusch and Moon (1984), and regression, such as in Jones and Mock (1984), were the methods more often used for performance assessments. Despite their popularity, they have some limitations. Each individual ratio examines only a part of the DMU activity, and therefore a comprehensive performance evaluation must be based on the analysis of several ratios. Therefore, it may be difficult to gain an overall view of performance, as the number of ratios that can be computed for each unit may be unmanageably large. The use of regression analysis for managerial purposes is widespread. However, it is based on average performance, whereas for benchmarking analysis the focus should be on the best observed practices. Given these limitations, alternative techniques have been proposed in the literature, such as Stochastic Frontier Approach (SFA) (Aigner *et al*, 1977) and DEA (Charnes *et al*, 1978). SFA imposes a parametric structure on the production technology and on the efficiency distribution, implying that all observations on the frontier must use the same technology. The DEA method uses the idea of assessing the efficiency of the DMUs without requiring the specification of a functional form for the production frontier. Therefore, it is defined by piecewise linear segments that connect the set of frontier observations, which correspond to the best performers. Although DEA has the advantage of imposing minimal assumptions on the shape of the production technology, it is a deterministic method, which does not account for random effects in the data. DEA has become the most widely used method for undertaking efficiency assessments. Some of the reasons that explain the preference for DEA in empirical contexts is that the technique is based on multi-input and multi-output frontier representations of the production technology, it does not require information on prices, and it can incorporate input and output variables measured in different scales. Furthermore, DEA results are easily obtained using linear programming.

The main focus of previous literature on retail productivity has been on the measurement and improvement of performance of companies from an industry (eg, Doutt, 1984; Good, 1984; Lusch and Moon, 1984; Ratchford and Brown, 1985; Ratchford, 2003), or retail stores from the same company (eg, Weitzel *et al*, 1989). Concerning the objectives of the analysis, several studies focused on labour productivity because the retailing activity is labour intensive, and therefore personnel expenditures are of great importance (eg, Ratchford and Brown, 1985; Athanassopoulos, 2004). Examples of studies that analysed other aspects that may influence store productivity, such as merchandise assortment, location, behavioural outcomes and environmental conditions, include Mahajan *et al* (1985), Donthu and Yoo (1998) and Thomas *et al* (1998).

The use of frontier techniques, such as DEA, has been recognised as a particularly appropriate method for performance assessments of stores within a company, such as in Thomas *et al* (1998), which assessed home furnishings and household items stores, and Grewal *et al* (1999) for stores of automobile parts. Concerning food-based outlets, few studies analysed the performance of multiple stores within the same organisation, such as Keh and Chu (2003), Barros and Alves (2004), Camanho *et al* (2009) and Vaz *et al* (2010), which analysed supermarkets from an organisation. DEA was also used to assess the efficiency of supermarket chains by Athanassopoulos and Ballantine (1995). The study described in this paper intends to define models based on DEA that fulfil the needs of retailing organisations management. The biggest challenge in performance assessments has been the unavailability of data, which hinders the development of robust models. Data availability often affects model development to an undesirable extent, where in fact modelling is done for the data available rather than the best formulation of the model.

3. Performance assessment methodology

This section describes the performance assessment method that enables setting store targets and comparing the performance between stores of different types. The DEA model used to set targets is described in the next section. The MI, which is used for performance comparisons between groups, is described in the following section. The index enables management to compare the efficiency spread in each group and the productivity differences between the best-practice frontiers of each group. The significance of the differences in group performance is tested using statistical hypothesis tests. Finally, we illustrate the integrated use of the MI and hypothesis tests with the analysis of four random samples generated by a Monte Carlo simulation.

3.1. DEA

DEA is a linear programming-based technique for measuring the relative efficiency of a fairly homogeneous set of DMUs in their use of multiple inputs to produce multiple outputs. It identifies a subset of efficient 'best-practice' DMUs and for the remaining DMUs, the magnitude of their inefficiency is derived by comparison to a frontier constructed from the 'best practices'. DEA derives a single summary measure of efficiency for each DMU. For the inefficient DMUs, DEA derives efficient input and output targets and a reference set (or peer group), corresponding to the subset of efficient DMUs to which they were directly compared.

The original DEA model proposed by Charnes *et al* (1978) assumed that all inputs and outputs can be varied at the discretion of managers. These may be called discretionary variables. However, often factors not subject to managerial control, called non-controllable variables, may also need to be considered in retailing performance assessments (Mahajan, 1991; Donthu and Yoo, 1998; Athanassopoulos, 2004). This is important to ensure fair comparisons, such that DMUs facing unfavourable conditions that they cannot influence are not penalised for producing less output or consuming more inputs than their peers. As the stores activity is critically affected by the external conditions in the catchment area, such as population density and number of competitors, it was important to account for their influence in the performance assessment described in this paper. This was accomplished by including in the input set both store resources and non-controllable factors reflecting environmental conditions. Note that since the DMUs assessment is output oriented, no input reductions are sought, and therefore both controllable and non-controllable factors can be included in the input constraints, with no differentiation between them in the formulation of the linear programming model.

In order to describe the formulation of the DEA model for an output oriented analysis, we define an input vector $\mathbf{x}=(x_1, \ldots, x_m) \in R_+^m$ used to produce an output vector $\mathbf{y}=(y_1, \ldots, y_s) \in R_+^s$ in a technology involving n production units. The efficiency of each DMU j_o is given by the reciprocal of the factor θ by which the outputs of the DMU j_o can be expanded:

$$\max\left\{ \begin{aligned} & h_{j_o} = \theta \,|\, x_{ij_o} \geqslant \sum_{j=1}^{n} \lambda_j x_{ij}, \quad i = 1, \ldots, m \\ & \theta y_{rj_o} \leqslant \sum_{j=1}^{n} \lambda_j y_{rj}, \quad r = 1, \ldots, s \\ & \lambda_j \geqslant 0, \quad \forall j \end{aligned} \right\} \tag{1}$$

Model (1) assesses the relative efficiency of the DMUs in the attainment of the output levels given the resources used and exogenous conditions. The measure of relative efficiency, given by $1/\theta^*$, is equal to 100% when the unit under assessment is efficient, whereas lower scores indicate the existence of inefficiencies. The efficient units are located in the frontier of the production possibility set identified by the DEA model. For the inefficient units, there is evidence that it is possible to obtain higher levels of outputs with the same or lower levels of the inputs currently used. For these units, it is also possible to obtain as by-products of the DEA efficiency assessment a set of targets for becoming efficient. The input and output targets for a DMU j_o under assessment are obtained as follows:

$$\begin{aligned} x_{ijo} &= x_{ij_o} - s_i^* = \sum_{j=1}^{n} \lambda_j^* x_{ij}, \\ y_{rjo} &= \theta_o^* y_{rj_o} + s_r^* = \sum_{j=1}^{n} \lambda_j^* y_{rj}. \end{aligned} \qquad (2)$$

The variables s_i^* and s_r^* are the slacks corresponding to the input i and output r constraints, respectively, obtained at the optimal solution to model (1). The benchmarks for the inefficient DMUs j_o are the units with values of $\lambda_j^* > 0$ in the optimal solution to model (1).

3.2. Malmquist index for group comparisons

The Malmquist index was introduced by Caves *et al* (1982) and developed further in the context of performance assessments by Färe *et al* (1994). The index is usually applied to the measurement of productivity change over time, and can be decomposed into an efficiency-change index and a technological-change index. Similarly, the performance index for group evaluation proposed by Camanho and Dyson (2006) can be decomposed in two effects: the relative positioning of the group frontiers that affect the productivity levels, and the efficiency spread within the groups. Thus, the comparison of different groups can be made through Malmquist indices adapted to a situation where different DMUs running under different programmes are compared, rather than the same unit in different periods of time.

Examples of this type of application can be found in Berg *et al* (1993) and Pastor *et al* (1997) in the context of comparisons between banks from different countries. The banks were first assessed in relation to their own country frontier and then the frontiers from different countries were compared using a Malmquist index. The Malmquist index used by these authors made use of an average DMU (bank) for the frontier comparisons. Each variable of the average DMU corresponded to the mean of that variable observed in all DMUs. These indices were based on the base period version of the Malmquist index introduced by Berg *et al* (1992). More recently, Camanho and Dyson (2006) proposed the use of Malmquist indices to compare group frontiers without the need to specify an average DMU. Instead, information regarding all DMUs is used in the Malmquist index computation.

As the index developed in Camanho and Dyson (2006) is used in the empirical part of this paper, we will describe it in more detail. The index is based on radial measures defined by distance functions. Camanho and Dyson (2006) described the input oriented version of the MI, whereas this paper uses an output-oriented index, which is consistent with the objectives of the retailing stores analysed in the empirical section. The output distance function is equal to the efficiency score estimated by model (1), which is $1/\theta^*$. Considering δ_A DMUs in group A, which use the inputs $x^A \in \mathbb{R}_+^m$ to obtain the outputs $y^A \in \mathbb{R}_+^s$, and δ_B DMUs in group B, which use the inputs $x^B \in \mathbb{R}_+^m$ to obtain the outputs $y^B \in \mathbb{R}_+^s$. The DMUs $j = 1, \ldots, \delta_A$ from group A are represented by their input-output vector as $(x_j^A,\ y_j^A)$. Let $D^A(x_j^A,\ y_j^A)$ be the distance function of DMU j belonging to group A when assessed in relation to technology A (defined by the DMUs belonging to group A), and $D^B(x_j^A,\ y_j^A)$ the distance function of the same unit assessed in relation to technology B (defined by the DMUs belonging to group B). For DMUs in group B, we can define similar measures. As the focus of this paper is the performance comparison of two groups of DMUs, we use the Malmquist index proposed by Camanho and Dyson (2006), which aggregates the distance measures obtained for all DMUs in each group through a geometric average, taking the form shown in (3). This aggregation enables to compare globally the groups' performance.

$$I^{AB} = \left[\frac{\left[\prod_{j=1}^{\delta_A} D^A(x_j^A, y_j^A)\right]^{\frac{1}{\delta_A}}}{\left[\prod_{j=1}^{\delta_B} D^A(x_j^B, y_j^B)\right]^{\frac{1}{\delta_B}}} \times \frac{\left[\prod_{j=1}^{\delta_A} D^B(x_j^A, y_j^A)\right]^{\frac{1}{\delta_A}}}{\left[\prod_{j=1}^{\delta_B} D^B(x_j^B, y_j^B)\right]^{\frac{1}{\delta_B}}} \right]^{\frac{1}{2}} \qquad (3)$$

In terms of interpretation, a score of $I^{AB} > 1$ indicates better performance in group A than in group B.

This index can be decomposed in the usual way in two components ($I^{AB} = IE^{AB} \times IF^{AB}$), following the approach by Färe *et al* (1994). One of the components compares the efficiency spread within the groups (IE^{AB}), and the other component compares the relative position of the group frontiers (IF^{AB}). This decomposition means that the sources of better performance can be associated with two factors: less dispersion in the efficiency scores of the DMUs within the group, and/or better productivity associated to the group frontier.

The index IE^{AB} (4) compares the efficiency spread within the groups. A value of $IE^{AB} > 1$ means that the efficiency spread is smaller in DMUs from group A than in those from group B.

$$IE^{AB} = \frac{\left[\prod_{j=1}^{\delta_A} D^A(x_j^A, y_j^A)\right]^{\frac{1}{\delta_A}}}{\left[\prod_{j=1}^{\delta_B} D^B(x_j^B, y_j^B)\right]^{\frac{1}{\delta_B}}} \tag{4}$$

The index IF^{AB} (5) compares the relative position of the group frontiers by measuring the distance between the two frontiers. This index is obtained as the geometric mean of two components (ratios). The first component is the geometric mean of the distances between the frontiers A and B, when assessed for the DMUs in group A. The second component is calculated in a similar way for the DMUs in group B.

$$IF^{AB} = \left[\left(\prod_{j=1}^{\delta_A} \frac{D^B(x_j^A, y_j^A)}{D^A(x_j^A, y_j^A)}\right)^{\frac{1}{\delta_A}} \times \left(\prod_{j=1}^{\delta_B} \frac{D^B(x_j^B, y_j^B)}{D^A(x_j^B, y_j^B)}\right)^{\frac{1}{\delta_B}}\right]^{\frac{1}{2}} \tag{5}$$

If the ratios are higher than 1 for all DMUs of the two groups, then the frontier of group A envelops the frontier of group B. This implies that the frontiers do not cross over and $IF^{AB} > 1$. If there is at least one DMU with a ratio < 1 and another DMU with a ratio > 1, the frontiers cross over. Note that the value of IF^{AB} results from aggregating ratios obtained for individual DMUs, and therefore the value of the index IF^{AB} is not enough to characterise the relative position of the group frontiers. This can only be inferred by the analysis of individual ratios prior to their aggregation in the Malmquist index.

3.2.1. Complementing a Malmquist index analysis with hypothesis tests. The objective of this section is to define a procedure for verifying, if the differences in performance between two groups evaluated using a MI are statistically significant. This enables the identification of the significant effects that make one group outperform the other. The choice of the adequate statistical tests for this purpose and the description of the procedure combining the use of the Malmquist index with statistical analysis is an important methodological contribution of this paper.

The procedure proposed consists of two main steps:

(1) In the first stage, the indices I^{AB}, IE^{AB} and IF^{AB} are calculated.
(2) In the second stage, hypothesis tests are used to verify if the differences between groups, in terms of efficiency and productivity levels captured by the indices IE^{AB} and IF^{AB}, respectively, are statistically significant.

To test, if the relative position of the group frontiers (evaluated by the index IF^{AB}) is statistically different, we used the Kolmogorov–Smirnov test (K–S test), following the proposal of Banker (1996) regarding the definition of appropriate statistical tests in the context of a DEA analysis. The K–S test assesses that if two independent samples are drawn from two similar populations (or populations with the same distribution). Note that we could have selected other non-parametric tests such as M–W or Median. For comparisons involving more than two groups, it is recommended to use the Kruskal–Wallis test, which can be used for various independent samples. To compare the location of the group-specific frontiers, we calculated for the DMUs in each group the efficiency distributions with reference to the own group frontier and with reference to the other group frontier. As the DMUs used as reference to calculate the efficiency scores in each group are different, the samples derived are independent. Thus, for the DMUs in group A, the null hypothesis compares the distribution of the estimates of $D^A(x_j^A, y_j^A)$, denoted by A_{own_j} , with the distribution of the estimates of D^B (x_j^A, y_j^A) denoted by A_{other_j} (H_o: *Dist.Effic.* A_{own_j} = *Dist.Effic.* A_{other_j}). The same procedure was used for DMUs in group B. This strategy was also used by Cummins *et al* (1999) to identify the most productive frontier. Table 1 summarises the hypothesis tests used to compare the relative position of the frontiers.

In terms of the results that may be obtained in the two K–S tests described in Table 1, the following situations could occur: (i) H_o is rejected in the two groups, meaning that the distance between the group frontiers is large and statistically significant; (ii) H_o is not rejected in any test, meaning that the frontiers of the groups are similar; (iii) H_o is only rejected in one group, meaning that for some input and output mixes, mainly observed in the group where H_o was not rejected, the frontiers are close to each other, whereas for the other input and output mixes the frontiers are further apart.

The index IE^{AB} evaluates the difference between the efficiency spreads within the groups. We propose using the K–S test to analyse the statistical significance of this difference. This involves comparing the distribution of the dispersion in the efficiency scores for the DMUs in group A (denoted by A_{own_j}) with the dispersion in the efficiency scores for the DMUs in the other group (denoted by B_{own_j}). These samples are also independent. This test is summarised in Table 2. If the null hypothesis is rejected, it can be concluded that efficiency spreads within the groups are significantly different.

Table 1 Hypothesis tests to compare the relative position of the frontiers (IF^{AB})

Test	H_o
K–S test for the DMUs in group A	H_o: *Dist.Effic.* A_{own_j} = *Dist.Effic.* A_{other_j}
K–S test for the DMUs in group B	H_o: *Dist.Effic.* B_{own_j} = *Dist.Effic.* B_{other_j}

3.2.2. Illustration of the method using a Monte-Carlo simulation. The objective of this section is to illustrate the integrated use of the methods described in the previous sections, which combine a MI with hypothesis tests. For that we generate four random samples by a Monte Carlo simulation. The four examples illustrate different scenarios concerning the relative position of the frontiers and the efficiency spreads within groups. Case I represents similar frontiers, but different efficiency spreads within groups. Cases II and III represent different frontiers, but similar efficiency spreads within groups. In case II one frontier envelops the other, and in case III the frontiers crossover. Case IV is identical to case III, but the number of DMUs in group A is half the number generated for case III, in order to investigate the influence of sample size. The data sets consisted of 200 DMUs for each group (except for group A in case IV, which has only 100 DMUs). We assume that each unit uses two inputs (x_1 and x_2) to produce one output (y), as shown in Figure 1. To allow a graphical representation of the random samples, the inputs were normalised by the value of the output (x_1/y and x_2/y). The technology was assumed to have constant returns to scale.

Table 2 K–S test to compare the efficiency spreads within the groups (IE^{AB})

Test	H_o
K–S test	*Ho: Dist.Effic.* A_{own_j} = *Dist.Effic.* B_{own_j}

The inputs were independently generated from uniform distributions and the efficient output levels were obtained using a known underlying technology (Cobb–Douglas), as described in Table 3. The parameters of each function were arbitrarily decided upon to enable testing the procedure developed in this paper.

We generated 20% of the DMUs to be Farrell efficient. The remaining DMUs were allocated an inefficiency component from a half-normal distribution, as defined in Table 4. A random noise term was also added with a standard deviation of 0.005.

The results of the procedure developed in this paper, including the Malmquist index and its components, as well as the hypothesis tests are reported in Table 5 for the four cases considered. The hypothesis tests were obtained using SPSS, at a significance level of 5%. For each case, we report the values of the indices I^{AB}, IF^{AB} and IE^{AB}, the p value (p) and the conclusion reached.

For case I, the hypothesis tests indicated that the only significant difference concerns the efficiency spreads within groups. This implies that the worst performance of

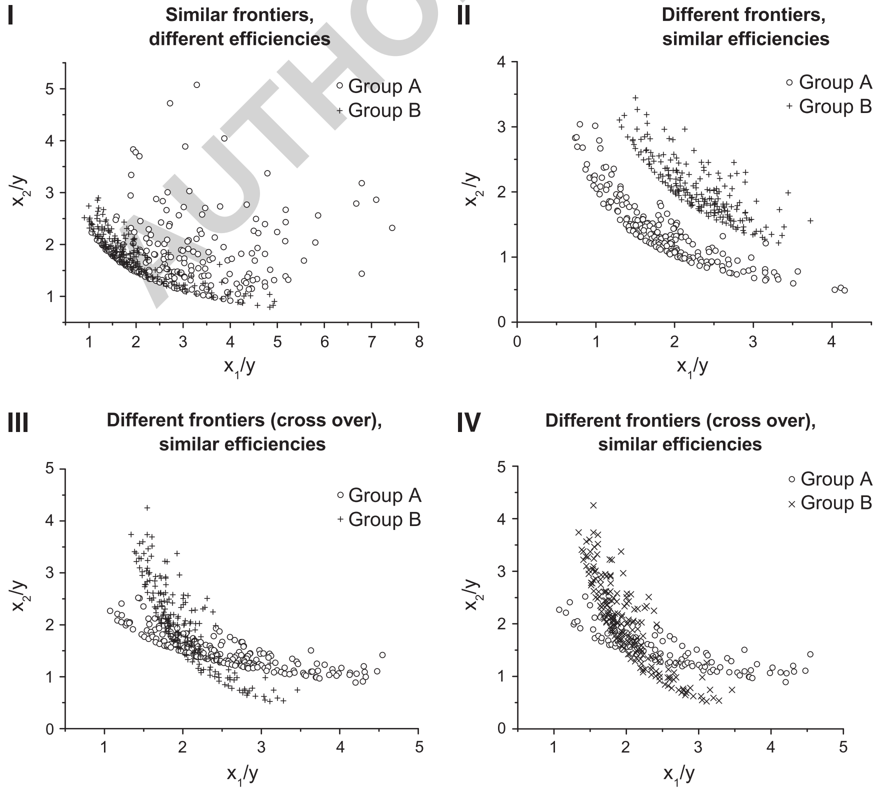

Figure 1 Illustration of the samples generated for cases I, II, III and IV.

Table 3 Functional form of the efficient frontier for cases I–IV

Case	*Efficient frontier for group A*	*Efficient frontier for group B*
Case I	$yA=0.6\ x_1^{0.4}\ x_2^{0.6}$, $x_1\rightarrow U^{250,1080}$ and $x_2\rightarrow U^{200,600}$	$yB=0.6\ x_1^{0.405}\ x_2^{0.595}$, $x_1\rightarrow U^{310,800}$ and $x_2\rightarrow U^{100,900}$
Case II	$yA=0.7\ x_1^{0.5}\ x_2^{0.5}$, $x_1\rightarrow U^{100,760}$ and $x_2\rightarrow U^{80,550}$	$yB=0.5\ x_1^{0.5}\ x_2^{0.5}$, $x_1\rightarrow U^{310,1080}$ and $x_2\rightarrow U^{380,830}$
Case III	$yA=0.6\ x_1^{0.4}\ x_2^{0.6}$, $x_1\rightarrow U^{250,1080}$ and $x_2\rightarrow U^{200,600}$	$yB=0.55\ x^{0.7}\ x^{0.3}$, $x_1\rightarrow U^{310,800}$ and $x_2\rightarrow U^{100,900}$
Case IV	$yA=0.6\ x_1^{0.4}\ x_2^{0.6}$, $x_1\rightarrow U^{250,1080}$ and $x_2\rightarrow U^{200,600}$	$yB=0.55\ x_1^{0.7}\ x_2^{0.3}$, $x_1\rightarrow U^{310,800}$ and $x_2\rightarrow U^{100,900}$

Table 4 Inefficiency distributions for the samples generated

Case	*Group A's inefficiency*	*Group B's inefficiency*
Case I	Half-normal (0,0.4)	Half-normal (0,0.1)
Case II	Half-normal (0,0.1)	Half-normal (0,0.11)
Case III	Half-normal (0,0.1)	Half-normal (0,0.11)
Case IV	Half-normal (0,0.1)	Half-normal (0,0.11)

Table 5 Results of the Malmquist index and statistical tests for cases I–IV

	Efficiency comparison	*Frontier comparison*
Case I:	$IE^{AB}=0.834$	$IF^{AB}=0.996$
Similar frontiers	K–S test: $p\approx 0\Rightarrow H_o$ rejected	K–S tests: for group A: $p=0.987\Rightarrow H_o$ not rejected
Different efficiencies		for group B: $p=0.142\Rightarrow H_o$ not rejected
$I^{AB}=0.831$	Conclusion: within-group efficiencies are different	Conclusion: the frontiers are similar
Case II:	$IE^{AB}=0.996$	$IF^{AB}=1.424$
Different frontiers	K–S test: $p=0.711\Rightarrow H_o$ not rejected	K–S tests: for group A: $p\approx 0\Rightarrow H_o$ rejected
Similar efficiencies		for group B: $p\approx 0\Rightarrow H_o$ rejected
$I^{AB}=1.418$	Conclusion: the efficiencies are similar	Conclusion: the frontiers are different
Case III:	$IE^{AB}=0.995$	$IF^{AB}=0.994$
Different frontiers	K–S test: $p=0.393\Rightarrow H_o$ not rejected	K–S tests: for group A: $p\approx 0\Rightarrow H_o$ rejected
Similar efficiencies		for group B: $p\approx 0\Rightarrow H_o$ rejected
$I^{AB}=0.990$	Conclusion: the efficiencies are similar	Conclusion: the frontiers are different
Case IV:	$IE^{AB}=0.990$	$IF^{AB}=0.990$
Different frontiers	K–S test: $p=0.176=H_o$ not rejected	K–S tests: for group A: $p\approx 0\Rightarrow H_o$ rejected
Similar efficiencies		for group B: $p\approx 0\Rightarrow H_o$ rejected
$I^{AB}=0.980$	Conclusion: the efficiencies are similar	Conclusion: the frontiers are different

group A compared with group B revealed by the index $I^{AB}<1$ (equal to 0.831) is because of the larger efficiency spread associated to this group ($IE^{AB}=0.834$).

For case II, the tests indicated that only the difference in the relative position of the frontiers is statistically significant. This implies that the best performance of group A revealed by $I^{AB}=1.418$ is because of the highest productivity levels associated to the frontier of this group ($IF^{AB}=1.424$).

For case III, both indices IE^{AB} and IF^{AB} are very close to 1, meaning that the efficiency spread within the groups is similar, and, on average, the productivity of the frontiers is the same. However, the hypothesis tests revealed that the relative position of the frontiers is different at a statistically significant level. As IF^{AB} is close to 1, this implies that none of the groups dominates the other in productivity terms for all the input-output mixes observed in the sample. However, some of the ratios underlying the calculation of the index IF^{AB} are significantly above 1, whereas others are below 1, meaning that the frontiers cross over, and in fact there are significant differences between the location of the frontiers. In order to identify the most productive frontier for some regions of the production possibility set, we have to analyse the individual ratios underlying the calculation of the index IF^{AB}. The advantage of using the procedure proposed in this paper is to highlight situations where differences between groups are significant, but where the MI, without the association with hypothesis tests, would lead to a misleading conclusion that the groups have similar frontiers. The correct conclusion is that the average frontier productivity is the same, but the frontiers are different.

Case IV is similar to case III and only intends to explore if differences in sample size would bias the results. As the analysis of case IV leads to similar conclusions as those obtained for case III, concerning the comparison of efficiency spreads and frontier productivity, we conclude that the hypothesis tests can lead to the correct conclusion even for analysis involving unequally sized samples.

The methodology developed in this paper, which associates the use of Malmquist indices with statistical tests, has successfully reached correct conclusions concerning the differences in frontier productivity and efficiency spreads between groups for the illustrative examples considered.

4. Empirical analysis of retail store efficiency and productivity

4.1. Context: The Portuguese retail sector

In the 1980s, the Portuguese retail sector was characterised by the existence of several small stores. The level of competition was low, and most companies had reduced capacity to innovate and limited power to negotiate with suppliers. With Portugal's entrance to the European Union (EU) in 1986, the sector embarked on transformations caused, among other factors, by increased competition, changes in consumer behaviour and the improvement of the country's socio-economic conditions. The increase in families' income was accompanied by an increase in indebtness of families, particularly incentivised by the reduction of interest rates and inflation. This caused a change in the families consumption pattern, with higher amounts spent on accommodation, transports, communication and consumer electronics, and a reduction of consumption associated with basic needs, such as food and clothes. These factors significantly modified consumer behaviour and affected the strategy of retailing companies. Consumer behaviour is characterised by making multipurpose shopping trips, combining purchases for different product categories and reducing the number of trips at a particular time period (Leszczyc *et al*, 2004). This is derived by the increased need for shoppers to optimise their time spent shopping because demands of every day professional and personal life have increased for most shoppers. Retailers have responded to this need by providing a wide assortment of products allowing consumers to combine purchases in multiple product categories.

The Portuguese retail sector is nowadays a mature sector, whose importance to the Portuguese economy increased significantly in the past few years, rising from a volume of business of €2634 million in 1988 to €10 710 million in 2005 (according to the retailing statistics compiled by the Nielsen company). The sector is dominated by four commercial groups, two Portuguese (Sonae, Jerónimo Martins) and two French (Intermarché, Auchan). Competition is high, and has been intensified by the entrance to the market of discount stores. The balance between the number of traditional grocery stores and the number of stores with modern formats (ie, supermarkets and hypermarkets with large sales areas) parallels the levels observed in Europe (according to the Portuguese Association of Retailing Companies – APED, the average market share of stores with modern formats in Portugal is 78% and in Europe is 86%).

In this highly competitive context, the downward pressure on sales margins demands additional efforts to rationalise processes and increase operations control, as well as to improve customer services and to maintain a loyal relationship with costumers. This makes efficiency assessment and improvement a key objective of retail organisations.

4.2. The organisation used as case study

The organisation used as case study has two main store types, hypermarkets and supermarkets, which constitute a chain operating in Portugal. These groups differ in the stores' sales area and in the market conditions of stores' catchment area. The hypermarkets are located in large urban areas, whereas the supermarkets are located in smaller urban areas. The hypermarkets sales area ranges between 4120 m^2 and 18 670 m^2, whereas the supermarkets area is about 2850 m^2. The layout of the stores analysed is organised in five sections: grocery (includes non-perishable food and drinks), perishables (includes meat, fish, fruit, vegetables and bread), textiles (includes footwear and clothes for men, women and children), household goods (includes cleaning products and books) and household appliances (includes hardware, audio, video and computers). We analysed a sample consisting of 18 hypermarkets and 18 supermarkets. The stores were selected by the managers of the company, to ensure homogeneity in the sample analysed and relevance of the results for the management of the store network, particularly for the planning and control department of the company. The data collected included all hypermarkets of the organisation at the time of this study. All stores had a layout with the household appliances section located in a separate area next to the main store.

The activity of each store is defined both by central management and local store management. Central management is in charge of the negotiation of contracts with suppliers, the definition of promotional policies and the selection of products and establishment of prices. The main decisions that are under the responsibility of local managers concern the organisation of the promotional events, the layout of the products in the shelves, the personnel recruitment and the store image. It is also a responsibility of the store to control the stocks and number of products spoiled, and monitor the adjustment of prices to the competitors and product ranges to the customer needs. The DEA model enables the modeller/management to estimate store efficiency and to set targets for inefficient stores. This efficiency measure reflects the ability of local management to operate close to the group-specific best-practice frontier.

4.3. Model specification

The DEA model used to assess the two groups of stores, hypermarkets and supermarkets, was output oriented with constant returns to scale, as described in (1). Thus, the store efficiency score, $1/\theta^*$, includes all the inefficiency sources related to scale size and resource under-utilisation. Scale efficiency reflects the inefficiency because of store size whereas pure technical efficiency reflects inefficient operation of the store.

In order to model the store activity, the input-output set should cover the full range of resources used and capture the outputs that are relevant for the objectives of the analysis. Good (1984) proposes a list of possible measures of retail outputs and inputs. Outputs are usually measured by the number of transactions, physical units sold, value added and sales value. Inputs are measured as the hours of labour employed, number of employees, wages, area of the store, inventory and advertising cost. According to Mahajan (1991), Donthu and Yoo (1998) and Athanassopoulos (2004), the inputs and outputs for retail productivity assessments should include controllable and uncontrollable factors (such as competitive conditions, population and per capita income). Although these factors are not subject to managerial control, they also need to be considered in the performance assessments to ensure fair comparisons. Thus, the DEA model should include the resources used, the outputs achieved and the uncontrollable factors, which are relevant to contextualise the assessment. Next, we describe the main input and output measures used in previous DEA studies of retailing services. The main inputs used were floor area, number of employees, stock and operational expenses (see Athanassopoulos and Ballantine, 1995; Thomas *et al*, 1998; Grewal *et al*, 1999; Keh and Chu, 2003; Barros and Alves, 2004; Camanho *et al*, 2009). The outputs can include sales value (in Athanassopoulos and Ballantine, 1995; Grewal *et al*, 1999; Keh and Chu, 2003; Camanho *et al*, 2009), sales value and profit (in Thomas *et al*, 1998) or sales value and operational results (in Barros and Alves, 2004).

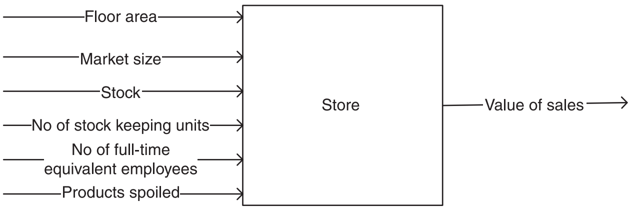
Floor area
Market size
Stock
No of stock keeping units
No of full-time equivalent employees
Products spoiled
Store
Value of sales

Figure 2 Inputs and output of store.

The inputs and outputs defined for evaluating store performance are described in Figure 2. The resources included in the DEA model were the floor area, the value of the products in stock, the number of stock keeping units (ie, the number of different products available in the store), the value of the products stolen or spoiled and the number of full-time equivalent employees. The input market size reflects the environmental conditions faced by the stores, as more favourable market conditions promote higher sales.

The output of the model is the total value of store sales. This enables to measure the ability of each store to maximise the sales by using its resources, taking into account the environmental conditions. Sales maximisation is the main objective of store managers, as the performance of the stores is assessed by the planning and control department of the retailing organisation based on a comparison of the store total sales with the target specified annually by the administration. Although this is a crude measure of the activity of each store, that could be further refined by separating the sales by individual store sections, this is outside the scope of this paper. A detailed analysis of performance of store sections and optimisation of resource allocation within the stores is available in Vaz *et al* (2010). Note that central management is in charge of the negotiation of contracts with suppliers, the definition of promotional policies and the selection of products and establishment of prices for each store of the chain. Thus, the stores in each group tend to sell the products at the same price. There is an exception for perishables products (because of their rapid deterioration) and for some products that have the highest visibility to customers, as shoppers tend to memorise their prices. These are determined taken into account the prices observed in competitors located in the surrounding area of a given store. Therefore, each store does not have direct responsibility in maximizing profit but rather in maximizing the total sales value. This implies that the targets defined by the central management for each store are based on sales value.

The floor area of the store represents its size, which has a direct influence on the volume of sales. According to a study undertaken by the company used as case study, a store with a larger floor area is more appealing to customers. The study concluded that the customer has the perception that a larger store has everything he/she needs. Thus, the floor area has a favourable impact on sales. Considering two stores with identical sales, the store that achieves these results with less floor area should be evaluated as being more efficient than the other. As argued by Desmet and Renaudin (1998) and Campo and Gijsbrechts (2004), floor area is considered the most important resource for retailers, although in most cases it is not controllable by store managers, at least in the short run. All the stores analysed in this study have similar equipments (eg, shelves, freezing equipment, cutting and packing machines), and therefore it was not considered necessary to specify a different input to represent this factor of production.

The stock is the value of the products that each store has available to sell. Considering two stores with identical sales, the store that achieves this results with less stock should be evaluated as being more efficient than the other. The number of stock keeping units represents the diversity of products available in the store. Thus, a store with a larger number of stock keeping units can satisfy the customer needs to a greater extent. Our model intends to assess the capacity of the store to maximise sales taking into account the variety of products sold and the value invested in stock, so both variables were included as inputs of the model, despite the high correlation between them.

The products spoiled relates to the amount lost with products stolen, damaged, spoiled or whose validity expired. Although this variable is a result of the activity of the store, it is an undesirable output that the store wants to minimise. There are several alternatives for including this type of data in the DEA models (see Dyson *et al*, 2001). To make this variable isotonic, it can be included in the model as an input, it can be deducted from a large constant or it can be inverted. The last two alternatives modify the measurement scale, which can make the interpretation of the results difficult. Thus, including the undesirable output as an input of the model was considered the best option for the analysis reported in this paper. As a result, stores with higher values of products spoiled are penalised in the DEA assessment.

We measured labour by the number of full-time equivalent employees. The input set does not include the cost of sales, because each store does not have responsibility in negotiating the price of products provided by suppliers, as previously explained.

It was also included in the input set the variable market size to characterise the demographic and competitive conditions of the store catchment area. The company analysed considers that the population and the competition are the most critical external factors that influence store activity. The population density, which has a positive impact on sales, is measured by the inhabitants living in municipalities within half-hour travel time from the store. The travel time is calculated considering an average speed of 50km/h. Conversely, competition has a negative contribution to sales. This variable is measured by the floor space of competitive stores within half-hour travel time from the store. Therefore, the variable used for representing market size is the population in the catchment area, adjusted by the floor space of the competitive stores, measured by the number of inhabitants per m^2 of competitors floor space.

Table 6 shows the summary statistics of the inputs and output of the 36 stores analysed. Table 6 shows that in hypermarkets the standard deviation of all variables is quite high relative to the mean, indicating a considerable amount of diversity in this type of stores.

5. Results and discussion

5.1. Target setting

The targets defined for the stores are determined based on internal benchmarking. This implies that each store is compared with other stores within the same group. The summary of the technical efficiency results obtained in each

Table 6 Mean and standard deviation values for the inputs and output of the hypermarkets and supermarkets

	Hypermarkets		*Supermarkets*	
	Mean	*Standard deviation*	*Mean*	*Standard deviation*
Inputs				
Floor area of the store (m^2)	8576	3601	2837	15
Market size	15	8	11	5
Stock of the store (euros)	5 779 444	2 258 766	1 658 737	158 402
No of stock keeping units	56 534	12 024	28 722	2141
Number of full-time equivalent employees	343	156	71	16
Products spoiled of the store (euros)	1 064 092	648 770	267 417	105 715
Output				
Sales of the store (euros)	83 553 352	39 090 745	16 079 548	3 953 074

group using the formulation shown in model (1) are presented in Table 7. The average technical efficiency values for the hypermarkets and supermarkets are shown in Table 7.

As shown in Table 7, the high levels of efficiency observed in each group indicate that store performance is rather homogenous, meaning that the scope for efficiency improvements is not very large. The assessment shows that there are five efficient hypermarkets and seven efficient supermarkets. In practice, observing the best practices of the efficient stores may help the worst performing DMUs to improve their performance.

For each inefficient store in each group, we can define the targets for performance improvement. This is an empirical evidence that their performance can be improved. For example, the technical efficiency of hypermarket M68 is 74%. The original values of the inputs and output of store M68, the DEA targets (calculated by (2)) and the peers are presented in Table 8. The main peer of hypermarket M68 in the DEA assessment is store M10, with a λ value equal to 0.294. The contribution of the other peer is marginal (store M07 has a λ value equal to 0.048).

The results indicate that there are two hypermarkets (M31 and M70) and six supermarkets (L22, L57, L62, L64, L65, L66) with slack in the constraint relative to the input market size. This means that these inefficient stores do not take full advantage of all their market potential, so it may be advisable to intensify efforts to increase sales. This can be done by increasing the sales of actual customers or attracting new customers. In these cases, it may be required a readjustment of the resources available at the store in order to fully explore the market potential.

Next, we compare the efficiency spread within each group of stores and the differences in the productivity of the frontiers.

Table 7 Efficiency results

Technical efficiency	*Hypermarkets*	*Supermarkets*
No of efficient stores	5	7
Average efficiency (%)	91.9	95.0
Standard deviation (%)	7.8	5.6

5.2. Groups comparison

5.2.1. Efficiency spread within groups. The value obtained for the index IE^{HS} (4) relating to the comparison of efficiency spread between hypermarkets (group H) and supermarkets (group S) was 0.965 (see Figure 3). The results of the statistical tests reported in Table 9 show that there are not significant differences in efficiency spread within the groups (both groups are similar in terms of efficiency achievements). Note that a value of the index $\mathrm{IE}^{\mathrm{HS}} < 1$ indicates that the supermarkets are closer to their best-practice frontier than the hypermarkets.

As the DEA model used assumed CRS, the efficiency estimate includes both a component relating to pure technical efficiency and a component of scale efficiency.

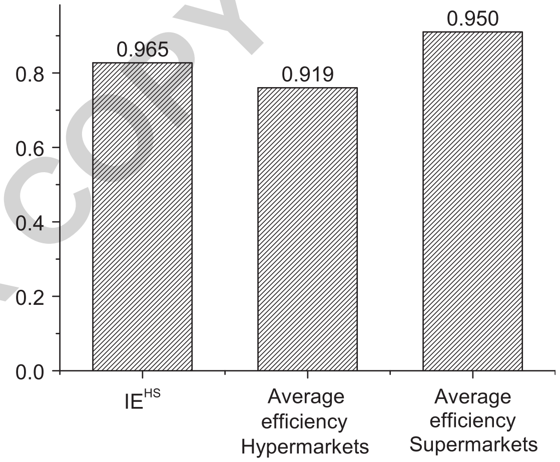

Figure 3 Value of index IE^{HS} and its components for the supermarkets and hypermarkets groups.

Table 9 Statistical tests to compare the efficiency spreads within groups

Efficiency comparison: $IE^{HS} = 0.965$
K–S test: $p = 0.7658 \Rightarrow H_o$ not rejected Conclusion: the efficiency spreads are similar

Table 8 Detailed analysis of hypermarket M68

	Observed in M68	*Target for M68*	*Peer store M07* ($\lambda = 0.048$)	*Peer store M10* ($\lambda = 0.294$)
Floor area of the store (m^2)	4440	2734	6044	8329
Stock of the store (euros)	2 970 537	2 133 490	4 065 348	6 606 010
Number of full-time equivalent employees	137	137	236	427
Products spoiled of the store (euros)	645 531	328 908	682 025	1 009 406
No of stock keeping units	40 200	19 467	49 520	58 254
Market size	12	12	38	33
Sales of the store (euros)	28 309 667	38 244 851	69 813 669	118 917 571

The scale efficiency is the ratio between the efficiency score achieved assuming CRS in model (1) and the efficiency score achieved assuming variable returns to scale, which requires including the constraint $\sum_{j=1}^{n}\lambda_j = 1$ in model (1). In order to see the impact of scale size on the efficiency estimates, we calculated the scale efficiency estimates for each group, and found that scale efficiency for the hypermarkets was, on average, 99.7% and for supermarkets was 100%. Therefore, we can conclude that the two groups are also similar in terms of scale efficiency achievements.

5.2.2. Relative position of the frontiers. The value of the index that compares the position of the frontiers, IF^{HS} (5), is equal to 1.639 (see Figure 4). The statistical tests indicated that the location of the frontiers is different at a statistically significant level (see Table 10). We can conclude that the productivity of the hypermarkets' frontier is greater than the productivity of the supermarkets' frontier.

In order to explore if the frontiers cross over, we analysed the ratios that estimate the distance between the frontier of the supermarkets and the frontier of the hypermarkets at the input-output mix of the DMUs observed in each of the groups $(D^S(X^i,Y^i)/D^H(X^i,Y^i), \forall i)$. The results obtained are reported in Table 11. For each store i, a ratio $D^S(X^i,Y^i)/D^H(X^i,Y^i) > 1$ means that the hypermarkets frontier has higher productivity than the supermarkets frontier for this input-output mix. The opposite occurs when $D^S(X^i,Y^i)/D^H(X^i,Y^i) < 1$.

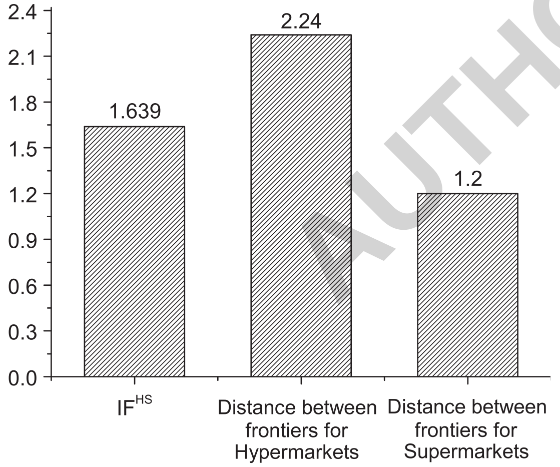

Figure 4 Value of the index IF^{HS} and the components relating to each group.

Table 10 Statistical tests to compare the relative position of the frontiers

Frontier comparison: $IF^{HS} = 1.639$
K–S tests: for Hypermarkets $p = 0.0 \Rightarrow H_o$ rejected
for Supermarkets $p = 0.0 \Rightarrow H_o$ rejected
Conclusion: the frontiers are different

Table 11 Analysis of the ratios that estimate the distance between frontiers $[D^S(X^i,Y^i)/D^H(X^i,Y^i)]$

Stores assessed	*Geometric mean*	*No. stores with ratio > 1*	*No. stores with ratio < 1*
All stores ($i = H, S$)	1.639	36	0
Hypermarkets ($i = H$)	2.240	18	0
Supermarkets ($i = S$)	1.200	18	0

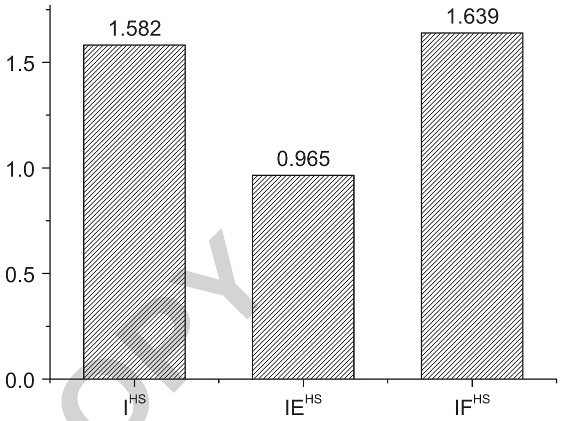

Figure 5 Values of the indices I^{HS}, IE^{HS} and IF^{HS}.

As the ratios were above 1 for all stores, we can conclude that the frontiers do not cross over, and the productivity of the hypermarkets' frontier is always greater than the productivity of the supermarkets, for all input-output mixes. This means that for the same input levels, the hypermarkets can obtain higher sales than the supermarkets. This is likely to occur, because the larger sales area of the hypermarkets enables having a larger diversity of products available, making these stores more attractive to customers. Hypermarkets are also located in large urban areas, with greater sales potential. Also, hypermarkets recruit knowledgeable and qualified staff and have a range of specialist services suited to the specific needs of each customer, which makes clients associate a premium quality of service to these stores. The quantitative analysis described in this paper enables to confirm that the benchmarking analysis should be done separately for each group of stores.

5.2.3. Overall group performance. The index reflecting overall group performance I^{HS} (3), that summarises the comparison of efficiency and productivity levels between the two store configurations, is equal to 1.582. Figure 5

Table 12 The profile of benchmark stores (hypermarkets)

	*Store M*03	*Store M*07	*Store M*10	*Store M*12	*Store M*69
Inputs					
Stock of the store (euros)	6 330 484	4 065 348	6 606 010	8 150 507	3 236 948
No of stock keeping units	62 536	49 520	58 254	64 232	41 015
Number of full-time equivalent employees	326	236	427	586	242
Products spoiled of the store (euros)	562 756	682 025	1 009 406	1 137 349	619 592
Area of the store (m^2)	8805	6044	8329	10 518	5065
Market size	10	38	33	12	9
Output					
Sales of the store (euros)	86 545 023	69 813 669	118 917 571	146 981 662	58 985 560

summarises the values of the overall Malmquist index and its components.

It can be concluded that the best performance of hypermarkets is because of the highest productivity of the group frontier. Nevertheless, there is still the potential to improve the efficiency levels of these stores. According to the DEA assessment, the hypermarket stores that can be considered benchmarks are M03, M07, M10, M12 and M69, and are characterised by having achieved the best performance standards both in terms of efficiency levels and frontier productivity. This means that these stores define the location of the efficient frontier and have the highest sales given the environmental conditions and resources used. The inputs and outputs of these stores are shown in Table 12.

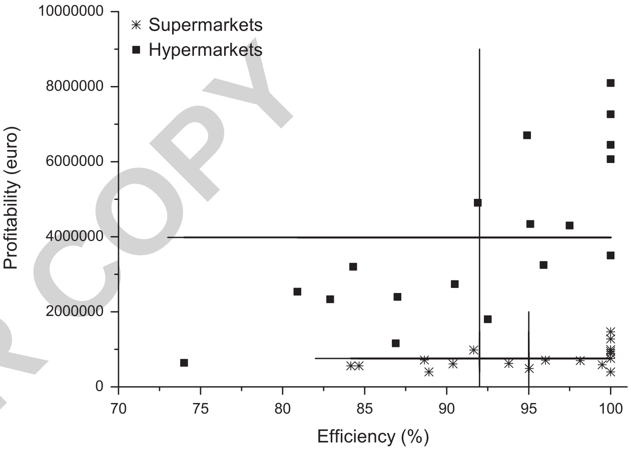

Figure 6 Stores' efficiency and profitability.

5.3. Profitability analysis

Figure 6 shows the relationship between technical efficiency and profitability of the stores in each group, following the framework proposed in Boussofiane *et al* (1991). In each group, the reference lines used are the average scores of efficiency and profitability. Globally, we can observe that hypermarkets are more profitable than supermarkets, which could be expected given the higher productivity of this type of stores verified in the previous section.

For each group, the contrast between efficiency and profitability intends to facilitate the appraisal of viability for individual stores. While efficiency assessments concentrate on the short run performance of individual stores, the confrontation of profitability and efficiency indicators enables analysing long-run viability.

The top right quadrant in Figure 6 corresponds to 'Star' stores as they are efficient in terms of sales and also have high profits. These stores should be used as the benchmarks of the organisation. These includes eight hypermarkets and six supermarkets.

The stores in bottom right quadrant have low profitability and high efficiency. These include three hypermarkets and five supermarkets. These stores are problematic as they are currently maximizing sales given the market conditions and resources used, but have difficulties in converting these sales into high profits.

The bottom left quadrant contains stores with poor performance in terms of efficiency and profitability. These include seven hypermarkets and six supermarkets. These stores should focus on attracting more customers in order to improve their sales, and eventually also increase profitability.

The top left quadrant contains stores with high profitability and low efficiency (called 'Sleepers'). Achieving high profits even without being efficient is an indication that the market conditions are favourable. There is only one supermarket located in this quadrant and should be a prime candidate for an efficiency improvement effort.

Regarding the benchmarks stores presented in Table 12, all hypermarkets are located in the 'Star' quadrant of Figure 6 with the exception of store M69, which has low profitability. This is probably explained by the less favourable exogenous conditions faced by this store.

6. Conclusions

This paper describes a performance assessment methodology that combines different management science methods to provide insights concerning the performance of stores. First, a DEA model is used to assess the stores performance and set appropriate targets. The targets defined for the stores are determined based on internal benchmarking as each store is compared with other stores of the same type (ie, hypermarkets or supermarkets). This procedure facilitates the identification of fair benchmarks for all stores, such that the targets provided can be supported by comparisons with similar stores. In practice, observing the best practices of the efficient stores may help the worst performing DMUs to improve their efficiency. Second, the integrated use of the MI and hypothesis tests enables to compare globally the performance of store groups, which requires characterizing their productivity levels. The MI is decomposed into sub-indices for comparing the efficiency spread in each group and the productivity differences between the best-practice frontiers of each group. The hypothesis tests are used to verify if the differences between groups captured by the sub-indices are statistically significant. The choice of the adequate statistical tests for this purpose and the description of the procedure combining the use of the MI with statistical analysis is one of the main methodological contribution of this paper. Finally, we illustrate the usefulness of the integrated use of the MI and hypothesis tests with the analysis of four random samples generated by a Monte Carlo simulation. These examples highlight situations where differences between groups are significant, although these differences could not be detected by a simple inspection of the Malmquist index score. The use of the MI associated to statistical hypothesis tests is essential to characterise the relative position of the group frontiers.

The methodology proposed can be used as an instrument for efficiency and productivity assessment of retailing stores, as illustrated in the analysis of a retail network of hypermarkets and supermarkets operating in Portugal. It was concluded that within each group of stores the performance is rather homogenous, meaning that the scope for efficiency improvements is not very large. The analysis also suggested that two hypermarkets and six supermarkets did not take advantage of all their market potential. These stores should try to increase sales, which may eventually require an adjustment to the level of resources used. This can involve, for example, the organisation of promotional events, the change of store layout, the renovation of the store image, a better recruitment of store operators, the improvement of stock management or the adjustment of prices and product ranges to fulfil customer needs. This should be undertaken by observing the best practices used by the benchmark stores identified in this research.

The comparison of performance between the groups provided evidence to support the conclusion that the hypermarkets have better performance than the supermarkets. Both store types are similar in terms of efficiency achievements in relation to their group-specific frontier. However, the hypermarkets' frontier is more productive than the supermarkets' frontier, which confirms that the benchmarking analysis of the stores should be done separately for the two groups.

The hypermarkets were also found to be more profitable than supermarkets, as could be expected given their higher productivity levels. There are eight hypermarkets and six supermarkets, which should be used as the benchmarks of the organisation ('Stars') as, within their group, they are efficient in terms of sales and also have high profitability. The practices observed in these stores should be disseminated to stores with poor performance in terms of efficiency and profitability.

References

Aigner D, Lovell CAK and Schmidt P (1977). Formulation and estimation of stochastic frontier production function models. *J Econometrics* **6**(1): 21–37.

Athanassopoulos AD (2004). Assessing the selling function in retailing. In: Cooper WW, Seiford LM and Zhu J (eds) *Handbook on Data Envelopment Analysis*. Kluwer Academic Publishers: Boston, pp 456–479.

Athanassopoulos AD and Ballantine JA (1995). Ratio and frontier analysis for assessing corporate performance: Evidence from the grocery industry in the UK. *J Opl Res Soc* **46**: 427–440.

Banker RD (1996). Hypothesis tests using data envelopment analysis. *J Prod Anal* **7**(2-3): 139–159.

Barros CP and Alves C (2004). An empirical analysis of productivity growth in a Portuguese retail chain using Malmquist productivity index. *J Retail Consum Serv* **11**: 269–278.

Berg SA, Førsund FR and Jansen ES (1992). Malmquist indices of productivity growth during the deregulation of Norwegian banking, 1980-89. *Scand J Econ* **94**(Supplement): S211–S228.

Berg SA, Førsund FR, Hjalmarsson L and Suominen M (1993). Banking efficiency in the Nordic countries. *J Bank Financ* **17**: 371–388.

Boussofiane A, Dyson RG and Thanassoulis E (1991). Applied data envelopment analysis. *Eur J Opl Res* **52**(1): 1–15.

Camanho AS and Dyson RG (2006). Data envelopment analysis and Malmquist indices for measuring group performance. *J Prod Anal* **26**(1): 35–49.

Camanho AS, Portela MC and Vaz CB (2009). Efficiency analysis accounting for internal and external non-discretionary factors. *Comput Opns Res* **36**(5): 1591–1601.

Campo K and Gijsbrechts E (2004). Should retailers adjust their micro-marketing strategies to type of outlet? An application to location-based store space allocation in limited and full-service grocery stores. *J Retail Consum Serv* **11**: 369–383.

Caves DW, Christensen LR and Diewert WE (1982). The economic theory of index numbers and the measurement of input, output and productivity. *Econometrica* **50**: 1393–1414.

Charnes A, Cooper WW and Rhodes E (1978). Measuring efficiency of decision-making units. *Eur J Opl Res* **2**(6): 429–444.

Cummins JD, Weiss MA and Zi H (1999). Organizational form and efficiency: The coexistence of stock and mutual property-liability insurers. *Mngt Sci* **45**(9): 1254–1269.

Desmet P and Renaudin V (1998). Estimation of product category sales responsiveness to allocated shelf space. *Int J Res Market* **15**(5): 443–457.

Donthu N and Yoo B (1998). Retail productivity assessment using data envelopment analysis. *J Retailing* **74**(1): 89–105.

Doutt JT (1984). Comparative productivity performance in fast-food retail distribution. *J Retailing* **60**(3): 98–106.

Dyson RG, Allen R, Camanho AS, Podinovski VV, Sarrico CS and Shale EA (2001). Pitfalls and protocols in DEA. *Eur J Opl Res* **132**: 245–259.

Färe R, Grosskopf S, Lindgren B and Roos P (1994). Productivity developments in swedish hospitals: A malmquist output index approach. In: Charnes A, Cooper WW, Lewin A and Seiford L (eds) *Data Envelopment Analysis: Theory, Methodology and Applications*. Kluwer Academic Publishers: Boston, pp 253–272.

Good WS (1984). Productivity in the retail grocery trade. *J Retailing* **60**(3): 81–97.

Grewal D, Levy M, Mehrotra A and Sharma A (1999). Planning merchandising decisions to account for regional and product assortment differences. *J Retailing* **75**(3): 405–424.

Jones K and Mock D (1984). Evaluating retail trade performance. In: Davies R and Rogers D (eds) *Store Location and Store Assessment Research*. John Wiley and Sons: New York.

Kamakura W, Lenartowicz T and Ratchford BT (1996). Productivity assessment of multiple retail outlets. *J Retailing* **72**(4): 333–356.

Keh HT and Chu S (2003). Retail productivity and scale economies at the firm level: A DEA approach. *Omega* **31**(2): 75–82.

Leszczyc PTLP, Sinha A and Sahgal A (2004). The effect of multi-purpose shopping on pricing and location strategy for grocery stores. *J Retailing* **80**(2): 85–99.

Lusch RF and Moon SY (1984). An exploratory analysis of the correlates of labor productivity in retailing. *J Retailing* **60**(3): 37–61.

Mahajan J (1991). A data envelopment analytic model for assessing the relative efficiency of the selling function. *Eur J Opl Res* **53**: 189–205.

Mahajan V, Sharma S and Srinivas D (1985). An application of portfolio analysis for identifying attractive retail locations. *J Retailing* **61**(4): 19–34.

Pastor JM, Perez F and Quesada J (1997). Efficiency analysis in banking firms: An international comparison. *Eur J Opl Res* **98**: 395–407.

Ratchford BT (2003). Has the productivity of retail food stores really declined? *J Retailing* **79**(3): 171–182.

Ratchford BT and Brown JR (1985). A study of productivity changes in food retailing. *Market Sci* **4**(4): 292–311.

Thomas RR, Barr RS, Cron WL and Slocum JW (1998). A process for evaluating retail store efficiency: a restricted DEA approach. *Int J Res Market* **15**(5): 487–503.

Vaz CB, Camanho AS and Guimaraes RC (2010). The assessment of retailing efficiency using network data envelopment analysis. *Ann Opns Res* **173**: 5–24.

Weitzel W, Schwarzkopf AB and Peach EB (1989). The influence of employee perceptions of customer service on retail store sales. *J Retailing* **65**(1): 27–39.

Received August 2010;
accepted April 2011 after one revision