The current issue and full text archive of this journal is available on Emerald Insight at:
https://www.emerald.com/insight/1741-0401.htm

Using Big Data to enhance data envelopment analysis of retail store productivity

Nicola Castellano
Department of Economics and Management, University of Pisa, Pisa, Italy
Roberto Del Gobbo
Department of Economics and Law, University of Macerata, Macerata, Italy, and
Lorenzo Leto
Department of Economics and Management, University of Pisa, Pisa, Italy

Received 8 April 2023
Revised 12 October 2023
Accepted 23 November 2023

Abstract

Purpose – The concept of productivity is central to performance management and decision-making, although it is complex and multifaceted. This paper aims to describe a methodology based on the use of Big Data in a cluster analysis combined with a data envelopment analysis (DEA) that provides accurate and reliable productivity measures in a large network of retailers.

Design/methodology/approach – The methodology is described using a case study of a leading kitchen furniture producer. More specifically, Big Data is used in a two-step analysis prior to the DEA to automatically cluster a large number of retailers into groups that are homogeneous in terms of structural and environmental factors and assess a within-the-group level of productivity of the retailers.

Findings – The proposed methodology helps reduce the heterogeneity among the units analysed, which is a major concern in DEA applications. The data-driven factorial and clustering technique allows for maximum within-group homogeneity and between-group heterogeneity by reducing subjective bias and dimensionality, which is embedded with the use of Big Data.

Practical implications – The use of Big Data in clustering applied to productivity analysis can provide managers with data-driven information about the structural and socio-economic characteristics of retailers' catchment areas, which is important in establishing potential productivity performance and optimizing resource allocation. The improved productivity indexes enable the setting of targets that are coherent with retailers' potential, which increases motivation and commitment.

Originality/value – This article proposes an innovative technique to enhance the accuracy of productivity measures through the use of Big Data clustering and DEA. To the best of the authors' knowledge, no attempts have been made to benefit from the use of Big Data in the literature on retail store productivity.

Keywords Productivity measures, Performance management, Big data, Tandem analysis, Data-driven clustering, Data envelopment analysis

Paper type Research paper

1. Introduction

Retail is one of the largest and most diversified industries (Kumar *et al.*, 2017). In 2021, global retail industry sales reached 26 trillion US dollars and are projected to exceed 30 trillion by 2024 (Statista, 2022). Generally, retail activities make a significant contribution to the national GDP of developed countries and serve as important drivers of economic growth in developing countries (Gandhi and Shankar, 2016).

© Nicola Castellano, Roberto Del Gobbo and Lorenzo Leto. Published by Emerald Publishing Limited. This article is published under the Creative Commons Attribution (CC BY 4.0) licence. Anyone may reproduce, distribute, translate and create derivative works of this article (for both commercial and non-commercial purposes), subject to full attribution to the original publication and authors. The full terms of this licence may be seen at http://creativecommons.org/licences/by/4.0/legalcode

International Journal of Productivity and Performance Management
Vol. 73 No. 11, 2024
pp. 213-242
Emerald Publishing Limited
1741-0401
DOI 10.1108/IJPPM-03-2023-0157

Although several factors related to technologies (i.e. self-checkouts, robotic automation and augmented reality) are expected to cause rapid market growth in the future, various limitations need to be considered, such as increasing operating costs and the consequent necessity to protect margins (The Business Research Company, 2021).

In response to the peculiar challenges and complexities that companies operating in the retail industry need to deal with, a broad research stream has developed specific measures to support decision-making and performance measurement. This article seeks to contribute to this research direction.

Among the indicators used to measure retail companies' performance, the traditional concept of productivity, calculated using an input/output ratio, is still central and relevant (Mishra and Ansari, 2013), as witnessed by the considerable number of studies on the subject published since the 1980s. Some of these studies have explicitly focused on developing context-specific techniques for calculating accurate and reliable productivity indices (Teng, 2014) by, for example, selecting the most significant input and/or output factors. Other studies have developed useful computational techniques to improve the precision and reliability of the results of productivity measurements (Günter and Gopp, 2021), such as linear regressions, translog cost functions, stochastic frontiers and data envelopment analysis (DEA). The DEA, which we employ in our study, has been extensively utilized in the retail sector (Vaz *et al.*, 2010) and various other domains (e.g. Majdi *et al.*, 2023). Many studies have endeavoured to enhance DEA by proposing innovative methodologies that often incorporate it with other techniques (Ebrahimnejad and Lotfi, 2012; Ebrahimnejad and Tavana, 2014; Nasseri *et al.*, 2018), thereby increasing its versatility, accuracy, reliability and applicability. For instance, Tavana *et al.* (2018) introduced an equivalence model that combined multi-objective linear programming problems with DEA, accommodating both desirable and undesirable factors, along with uncontrollable variables. Their approach facilitates interactive performance planning and allows decision-makers to strategize efficiency improvements within budget constraints. Likewise, Ebrahimnejad and Amani (2021) presented a novel DEA model that takes into account undesirable factors using triangular fuzzy numbers. Introducing the concepts of fuzzy ideal and anti-ideal decision-making units (DMUs), their approach enables comprehensive efficiency analysis, evaluates units from both optimistic and pessimistic perspectives and reduces computational complexity.

In brief, the productivity scores calculated through a DEA are the result of a benchmark on all the DMUs, where a score of 1 is assigned to the most efficient DMUs, while the score decreases as the level of productivity decreases, with a minimum of 0. From a managerial perspective, the lower is the score for a given DMU, the higher is the magnitude of the improvement requested to that DMU to reach the maximum level of productivity. Of course, this line of reasoning is valid under the assumption that all the DMUs are homogeneous and fully comparable in terms of utilization of inputs and production of outputs (Samoilenko and Osei-Bryson, 2008). When remarkable differences occur among the DMUs, which can be either internal or external exogenous factors, the results obtained through a DEA may lack accuracy. Heterogeneity exacerbates the distance between the least efficient DMUs and the most efficient ones, making the latter an unachievable target (Amirteimoori and Kordrostami, 2013). Furthermore, distorted performance evaluations may unfairly attribute responsibility to managers of less efficient DMUs when the lack of productivity is due to exogenous factors rather than poor managerial capacity (Hajiagha *et al.*, 2016; Li *et al.*, 2016).

Cluster analysis is widely employed in combination with DEA to mitigate heterogeneity. It enables the division of DMUs into sub-groups that exhibit homogeneity in one or more variables considered meaningful to the analyst, facilitating an accurate comparison of efficiency levels (Samoilenko and Osei-Bryson, 2008). In several cases, scholars focus on scale heterogeneity by referring to differences in the size of DMUs and consequently divide them into clusters homogeneous in terms of levels of input and output (Amirteimoori and

Kordrostami, 2013; Hajiagha *et al.*, 2016; Li *et al.*, 2016). In our study, cluster analysis is employed to address some of the significant sources of heterogeneity (Zarrin *et al.*, 2022). Specifically, we assume that structural characteristics and exogenous environmental variables play a substantial role in interpreting the variations in the productivity levels of the DMUs. On the other hand, we acknowledge that previous studies have warned about (Dyson *et al.*, 2001; Samoilenko and Osei-Bryson, 2010): 1) the difficulties related to identifying and measuring environmental variables; 2) the problem of dimensionality, which may determine a loss of discriminatory power in the analytical model when too many variables are included.

The paper aims to contribute to this body of literature by proposing a Big Data-based methodology that is valuable for utilizing exogenous environmental variables in cluster and DEA while simultaneously addressing the dimensionality problem.

The increasing development of Big Data allows the expansion of the company data ecosystem. It makes available innovative measurements for variables meaningful to support decision-making that usually are not represented in the company information system. Among the wide array of Big Data sources, in this study, we use human mobility data (HMD) and geographic information systems (GIS) to measure multiple aspects concerning the catchment area in a network of retail stores. To reduce the dimensionality, we propose a tandem analysis, which consists of a combination of factorial techniques (principal component analysis [PCA] and multiple correspondence analysis [MCA]) followed by an agglomerative hierarchical clustering (AHC). In coherence with a meta-frontier approach, the DEA is performed in two steps on the entire population and within the clusters. The comparison of the results provides evidence of the benefits of this kind of approach. In particular, the calculations of the retailers' productivity based on clustering showed higher levels of accuracy and reliability. This methodology can provide effective support in the management of control processes, specifically when it comes to setting achievable targets for retailers.

This technique was applied to a leading kitchen furniture producer that, as part of its development strategy, provided financial and organizational support to retailers for opening and renewing a massive number of stores. Therefore, it was essential for the company to measure the stores' productivity for performance evaluation purposes and to set adequate sales targets for retailers to improve the overall volume of sales and profitability.

The paper aims to contribute to the existing body of literature in multiple ways. While Big Data is widely used in clustering analysis, its applications in the retail sector remain limited. To date, there have been no attempts to combine Big Data cluster analysis with DEA. The peculiarities of the Big Data sources adopted in the analysis also contribute to the advancement of the studies. While GIS data is well-known in retail industry research, HMD represents a novelty. The use of GIS and HMD in combination provides the possibility of profiling the population that actually lives in a specific geographical area, rather than considering the population that is formally registered in the same area but may be residing elsewhere. Furthermore, Big Data provides a wide range of measures that enable data-driven detection of the environmental and structural characteristics that are significant for grouping retailers into homogeneous clusters to enhance the reliability of productivity analysis. Finally, the proposed tandem analysis is effective in reducing the dimensionality associated with the use of Big Data and, equally importantly, it can be easily adapted to analogous contexts.

This article is structured as follows: The second section opens with a brief literature review of DEA applications in retail store productivity. The section also contains a brief literature review of cluster analysis used to reduce the heterogeneity of DMUs, which is a major drawback in the DEA. The advantages of Big Data in cluster analysis are also discussed as a way to measure external exogenous variables. The third section presents the DEA methodology, while the fourth section is devoted to the empirical application of the

tandem analysis that combines Big Data clustering with DEA and to the discussion of the results. The last section presents the conclusions along with empirical implications, limitations and future research avenues.

2. Literature review

2.1 DEA adoption in retail store productivity

Scholars have been looking for ways to improve productivity measures for many years, as attested by numerous studies published on the subject (Bhat *et al.*, 2016; Islam and Syed Shazali, 2011). Although the importance of productivity for retail companies is widely acknowledged (Ingene, 1984), several aspects remain controversial, such as the determinants of retail performance and the coherence, accuracy and suitability of measures (Higón *et al.*, 2010), which has resulted in a long-running debate among academics and practitioners.

For example, some country-level studies have revealed differences among productivity drivers that could be relevant when developing suitable measures (Käpylä *et al.*, 2010; Griffith and Harmgart, 2005). In the United States, the productivity growth at the beginning of the 2000s was significantly driven by within-firm strategies (i.e. companies were opening new productive stores to replace non-productive ones; see Foster *et al.*, 2002), whereas in the United Kingdom, strategies of the same kind were associated with lower productivity (Haskel and Khawaja, 2003).

Another research strand has focused on developing effective measurement frameworks for productivity and efficiency at the store level, which are critical for assessing retailer competitiveness (Pestana Barros and Alves, 2003) and taking appropriate actions (Dubelaar *et al.*, 2002). Improvements in store productivity affect the overall performance of companies. Therefore, productivity measures at the store level provide greater and more detailed analytic support for decision-making than company-aggregated measures (Keh and Chu, 2003). As a result, productivity measures at the store level have been extensively discussed in the literature as a means of supporting a wide array of decision-making purposes, such as optimizing resource allocation (Yang, 2020), performance-based rewarding systems (Vyt and Cliquet, 2017), identification of technical inefficiencies and related solutions (Pestana Barros and Alves, 2003) and support for manager control and expansion strategies (Dubelaar *et al.*, 2002).

A productivity measure generally takes the structure of an output-to-input ratio (Käpylä *et al.*, 2010). Output indicators may include, for example, the number of units sold, sales revenues, footfall and the degree of customer satisfaction, while input indicators can involve employee numbers, labour costs, hours worked and store size (Donthu and Yoo, 1998).

Several issues may affect the measurement of productivity at the store level. First, a productivity ratio that includes only two variables, representing the outcome and the input, respectively, may be too simplistic to effectively represent the multiple facets of productivity (i.e. technical, economic and intangible, among others; see Kamakura *et al.*, 1996). Moreover, the comparison of productivity among stores located in different places may be biased due to significant differences in economic, social and demographic characteristics, which influence store performance in relative terms.

To address the complexities mentioned above, the literature has predominantly proposed statistical methods such as regression analysis (Donthu and Yoo, 1998), translog cost functions (Kamakura *et al.*, 1996), stochastic frontier (Barros, 2005; Gupta and Mittal, 2010; Gauri, 2013) and DEA.

DEA (Farrell, 1957; Charnes *et al.*, 1978) is a non-parametric technique that allows for great flexibility in calculating productivity and efficiency because multiple inputs and outputs of different nature (e.g. continuous, numerical and categorical; see Banker and Morey, 1986) can be summarized in the same measurement framework, including qualitative figures (Nyhan and Martin, 1999).

A further advantage of DEA is the ability to calculate optimal values through linear programming, as well as its easy adaptability to different contexts (Zhou and Xu, 2020). Given its peculiar characteristics, DEA provides effective support in comparing multiple DMUs whose performance and productivity are influenced by different environmental conditions.

In their seminal work, Charnes *et al.* (1985) developed the premises for applying DEA to calculate retail store productivity. Afterwards, a similar approach has been used in a wide array of cases. A summary of the literature on DEA adoption in the retail industry is summarized in Table 1.

Scholars over time have contributed to the literature by proposing approaches to using DEA for ranking purposes, aiming to assess the most efficient retailers competing in the same country (e.g. Sellers-Rubio and Ruiz, 2006; Perrigot and Pestana Barros, 2008; Mostafa, 2009; Gandhi and Shankar, 2014). In several other cases, the use of DEA has been proposed aiming to provide managers with information about how to enhance the retailers' efficiency within the same company, thus improving the company efficiency overall (Pestana Barros and Alves, 2003; Vaz *et al.*, 2010; Ko *et al.*, 2017; Vyt and Cliquet, 2017, Gong *et al.*, 2019; Huang *et al.*, 2019; Rouyendegh *et al.*, 2019).

DEA is often used in combination with other methodologies in order to widen the scope of the research. Among others, bootstrapped Tobit regression is used to assess the impacts of the different determinants of efficiency (Perrigot and Barros, 2008; Gandhi and Shankar, 2016; Ko *et al.*, 2017). Gong *et al.* (2019) adopt DEA with hierarchical regression analysis and nonlinear analysis to evaluate how efficiency may be improved through sustainability initiatives, while Rouyendegh *et al.* (2019) use Intuitionistic Fuzzy Technique for Order of Preference by Similarity to Ideal Solution technique to include some qualitative variables in the calculation of efficiency.

The heterogeneity of DMUs can represent a major drawback in the use of DEA. The retailers very likely operate in different conditions that may act as boosters or hinderers of efficiency. If such conditions are not considered somehow in calculating the efficiency scores, the comparison of all the DMUs in the search for the best performers may be altered. The heterogeneity bias becomes relevant when the number of DMUs is remarkably high, as in the case study described in the present paper. Consequently, in the following section, we summarize the literature on how heterogeneity is treated in DEA adoption.

2.2 DEA and the problem of heterogeneity

The robustness of the results obtained when using DEA depends on the homogeneity of the DMUs (Jiang *et al.*, 2020). A comparison of DMUs with dissimilar characteristics produces non-accurate and unreliable efficiency scores (Samoilenko and Osei-Bryson, 2008; Lozano-Vivas *et al.*, 2002). In the case of non-homogeneous DMUs, the distribution of productivity scores is sparse, with many DMUs reporting low scores (Zhu, 2022).

Heterogeneity may depend on the violation of the following three basic assumptions (Dyson *et al.*, 2001): (1) the DMUs produce comparable products or services (same outputs), (2) the same set of resources is available to all the DMUs (same inputs) and (3) the DMUs do not operate in the same environment. Samoilenko and Osey-Bryson (2010) propose that homogeneity within a set of DMUs must be maintained from both a *semantic* and a *scale* perspective. Semantic homogeneity requires that all DMUs within the set share a common meaning for the decision maker, while scale homogeneity pertains to the comparability of input and output levels across all DMUs in the sample.

Heterogeneity can be reduced by limiting the analysis to a sample of DMUs selected through criteria considered meaningful by the analyst (Ko *et al.*, 2017; Vyt and Cliquet, 2017), while clustering analysis is used to a larger extent to divide the entire set of DMUs into

Authors	Purpose	Method	Input variables	Output variables	Nr of DMUs
Pestana Barros and Alves (2003)	To analyse the efficiency of retail stores of a Portuguese multi-market hypermarket retailing chain	Output-oriented VRS** DEA	1. Nr of employees; 2. Cost of labour; 3. Absenteeism; 4. Area of outlets; 5. Nr of points of sale; 6. Age of the outlet; 7. Inventory; 8; Other costs	1. Revenues; 2. EBIT	47
Sellers-Rubio and Ruiz (2006)	To estimate the economic efficiency of Spanish supermarket chains	Traditional non-parametric input-oriented CRS* DEA	1. Nr of employees; 2. Nr of outlets in supermarket chain; 3. Capital invested	1. Revenues; 2. Profits	100
Perrigot and Barros (2008)	To analyse the efficiency of the French retailers in order to identify the best-practice reference enterprises. To determine the determinants of retailers' efficiency	Two-step procedure: DEA + Bootstrapped Tobit. Four DEA models are used: CRS*; VRS**; cross-efficiency; super-efficiency	1. Nr of employees; 2. Total assets; 3. Total costs	1. Revenues; 2. Profits	11 companies x 5 Years
Mostafa (2009)	To measure the relative efficiency of the US specialty retailers and food consumer	Output-oriented VRS** DEA	1. Nr of employees; 2. Total assets	1. Revenues; 2. Market value; 3. Earnings per share	45
Vaz *et al.* (2010)	To assess efficiency in stores selling different lines of product	Network DEA. Two-stage analysis at line-of-product ad store-level	1. Floor area. 2. Value of products in stock. 3. Nr of references. 4. Value of products spoiled	1. Revenues	70
Gandhi and Shankar (2014)	To find the "best in class" between Indian retailers. To analyse the pattern of efficiency change over time. To test impacts of environmental factors on efficiency of firms	Input-oriented DEA (CRS* and VRS**); Malmquist Productivity Index, Bootstrapped Tobit Regression	1. Cost of labour; 2. Total assets	1. Profits; 2. Sales	18 companies x 3 Years

Table 1.
Summary of literature on DEA adoption in retail industry

(*continued*)

Authors	Purpose	Method	Input variables	Output variables	Nr of DMUs
Ko *et al.* (2017)	To measure the efficiency of individual stores. To assess the factors that affect store efficiency	DEA + Bootstrapped Tobit Regression	1. Store size. 2. Nr of employees. 3. Nr of items. 4. Rental costs	1. Revenues. 2. Nr of customers	32
Vyt and Cliquet (2017)	To measure retail performance at store level by taking into account the stores' local market characteristics	Two-step procedure: output-oriented DEA + OLS regression of efficiency scores upon 8 local variables	1. Store size. 2. Nr of employees. 3. Product shelf space allocation	1. Revenues	38
Gong *et al.* (2019)	To evaluate the retailers' benefits on efficiency coming by sustainable operations. To evaluate under which internal conditions an increase of sustainable operations will determine likely an improvement in operational efficiency	Two-stage DEA (evolution of CRS* model); hierarchical regression analysis; non-linear analysis	Stage 1: Supply chain coordination (4 variables); Sustainability level (compliance, environmental, created sharing values) Outputs of stage 1 are inputs for stage 2	Stage 1: Cost competency (4 variables); Flexibility competency (3 variables); Social competency (4 variables); Environmental competency (4 variables) Stage 2: Business performance (sales growth; profits growth; market share growth; ROI)	124
Huang *et al.* (2019)	To evaluate the performance of the allocation process in the fashion industry	Multi-stage efficiency model based on dynamic network DEA (CRS*)	1. Initial allocation quantity; 2. Replenishment quantity	1. Sales quantity; 2. Inventory quantity	52
Rouyendegh *et al.* (2019)	To evaluate efficiency in retail industry by using both quantitative and qualitative data	Intuitionistic Fuzzy Technique for Order of Preference by Similarity to Ideal Solution (IF-TOPSIS) + CRS* DEA	1. Nr of employees; 2. Parking area for the customers; 3. Average number of customers per m^2 daily	1. Amount of money per customer trip per m^2 daily; 2–3. Flexibility and accessibility (qualitative variables)	21

(*continued*)

Table 1.

Authors	Purpose	Method	Input variables	Output variables	Nr of DMUs
Our contribution	To measure efficiency among a large number of retailers, by taking into account the heterogeneity in store characteristics and in the socio-demographic traits of their catchment area	Tandem Analysis: Data-driven factorial and clustering of DMUs + Output-oriented VRS** DEA	1. Average number of kitchen models presented in the store; 2. Total costs of store setup	1. Sales quantity; 2. Revenues	541

Note(s): *CRS = constant returns to scale. It is also known as CCR: Charnes, Cooper, Rhodes (1978). **VRS = variable returns to scale. It is also known as BCC: Banker *et al.*, (1984)

Table 1.

homogeneous sub-groups (Charnes and Cooper, 1980; Vyt and Cliquet, 2017). Other studies include some variables of environmental or exogenous nature that could be meaningful in discriminating the DMU efficiency. Such variables can be included in the DEA model (Dyson *et al.*, 2001; Gong *et al.*, 2019; Rouyendegh *et al.*, 2019) or during a subsequent step of the analysis, aiming to determine the factors that have major impacts on efficiency and correct the scores previously obtained by the DEA model (Haas and Murphy, 2003; Vyt and Cliquet, 2017).

The adoption of clustering analysis in combination with DEA is preferable from a managerial perspective because it enables an explicit identification of natural groups of DMUs (Samoilenko and Osei-Bryson, 2008), and it is helpful to managers since it provides an assessment of the DMUs *status quo* under an efficiency perspective. Furthermore, clustering used in combination with DEA provides a quantification of more realistic targets for non-efficient (or less efficient) DMUs. A summary of the literature on clustering analysis used in combination with DEA is shown in Table 2.

In the existing literature, various clustering techniques are utilized to subgroup DMUs including, among others, hierarchical methods (agglomerative or divisive), partitional techniques (such as k-means clustering) and fuzzy clustering. These methods differ in the algorithms used to assign a DMU to a specific cluster and the approach taken to determine the optimal number of clusters in which the population under investigation should be partitioned. In several instances, the number of clusters is predefined by the analyst as a requirement of the clustering technique, as is the case with k-means clustering, or it is determined through domain expert knowledge. In contrast, clustering techniques that do not require any assumption by the analyst are applied in a limited number of cases (Sharma and Yu, 2009; Zarrin *et al.*, 2022). Clustering based on *a priori* assumptions is suitable when DMUs can be categorized into a taxonomy commonly used in the context of analysis. Conversely, unsupervised clustering aims to generate intrinsically effective clusters according to the characteristics of the dataset being used (Samoilenko and Osei-Bryson, 2019).

Two alternative methods of combining clustering Analysis and DEA emerge in the literature. In the first approach, the DEA is performed on the entire dataset, and cluster analysis is used to classify the DMUs on the DEA results. Alternatively, the cluster analysis precedes the DEA, which is then performed only within the cluster to calculate for each DMU a score of efficiency that is only relative to the cluster it belongs to. The first approach is used

Table 2.
Summary of literature on clustering analysis used in combination with DEA

Authors	Context of analysis	Clustering method	Clustering variables	Nr. of DMUs	Nr of clusters	Method of combination with DEA
Samoilenco and Esei-Bryson (2008, 2010)	Countries in transition from centralized economies to market economies	Two-step approach based on k-means	Levels of DMUs inputs and outputs	18	2 clusters, through user-specified threshold and domain expert knowledge	Clustering is adopted to DEA results on the entire dataset
Sharma and Yu (2009)	Container terminals	Kohonen's Self Organizing Map (KSOM) preceded by a stratifying method	Levels of DMUs inputs	70	4 clusters through unsupervised clustering	Clustering is adopted to DEA results on the entire dataset
Amirteimoori and Kordrostami (2013)	Retail: Bank branches	Original method based on branches size	Size	64	3 clusters by domain expert knowledge	Clustering is adopted prior to DEA. DEA is used within clusters
Hajiagha *et al.* (2016)	Retail: Bank branches	Fuzzy c-means clustering	Levels of DMUs inputs and outputs	117	2 clusters set by the analyst	Clustering is adopted prior to DEA. DEA is used within clusters
Li *et al.* (2016)	Retail: Gas stations	Ward's hierarchical clustering	Levels of DMUs inputs and outputs	197	4 clusters set by the analyst	Clustering is adopted to DEA results on the entire dataset
Omrani *et al.* (2018)	Hospitals	Fuzzy c-means clustering	Environmental characteristics (population; GDP per capita)	288	5 clusters set by the analyst	Clustering is adopted prior to DEA. DEA is used within clusters
Costa *et al.* (2019)	Electricity energy distribution utilities	Spatial Bayesian clustering	Spatial location (assuming that geographically closer DMUs are homogeneous)	64	2 clusters as result of the analysis	Clustering is adopted to DEA results on the entire dataset

(*continued*)

Table 2.

Authors	Context of analysis	Clustering method	Clustering variables	Nr. of DMUs	Nr of clusters	Method of combination with DEA
Samoilenko and Osei-Bryson (2019)	Sub-Saharan African countries	Hybrid partitional/hierarchical approach	Economic development; socioeconomic impact of ICT; growth in productivity	27	3 clusters by domain expert knowledge	Clustering is adopted to DEA results on the entire dataset
Cinaroglu (2020)	Hospitals	K-means clustering	regional areas are clustered on welfare indicators	81	5 clusters through a combination of different factors	Clustering is adopted prior to DEA. DEA is used within clusters
Zarrin *et al.* (2022)	Hospitals	Self-Organizing Map-Artificial Neural Network	hospital's characteristics	1,124	3 clusters through unsupervised clustering	Clustering is adopted prior to DEA. DEA is used within clusters
Tsionas (2023)	Commercial banks	Convex non-parametric least squares	Commercial banks operating variables nd technical feasibility	285	3 clusters through Bayesian Model Averaging	DEA is used on the entire dataset. Clustering is adopted to DEA results
Our contribution	Retail: Kitchen furniture	Factorial techniques (PCA and MCA) on Big Data + Combined Agglomerative Hierarchical Clustering and k-means algorithm method	Structural characteristics of stores and socio-demographic characteristics of their catchment area	541	3 clusters as result of an agglomerative hierarchical clustering, based on factorial scores	Clustering is adopted prior to DEA. DEA is used within clusters

to divide the investigated set of DMUs into multiple reference subsets of homogeneous DMUs whose efficiency is calculated on the entire group of DMUs (Sharma and Yu, 2009; Li *et al.*, 2016; Costa *et al.*, 2019, only to name a few). The second is also called the meta-frontier approach, which assumes that the DMUs operate in contexts with different characteristics that can be, among others, of environmental, social, cultural or technological nature. Consequently, it becomes coherent to evaluate the efficiency of the DMU within the clusters, then identifying group frontiers (or local frontiers), while the maximum levels of efficiency calculated on all the DMUs form the meta-frontier (Yu and Chen, 2020).

The meta-frontier DEA approach is coherent in research embracing a managerial perspective, where the DEA scores are helpful in assessing the level of efficiency of the DMUs and setting achievable improvement targets, with reference to a maximum level of efficiency that is compatible with their specific characteristics and operating context, which can be remarkably different from other DMUs (Hajiagha *et al.*, 2016; Li *et al.*, 2016; Zarrin *et al.*, 2022). Following the meta-frontier approach, we calculate DEA scores both on the entire set of DMUs and within the clusters, so determining two efficiency scores for each DMU: namely the "pooled" score and the "separate" (or "within") score, that are referred to the meta-frontier and to the group-frontier respectively. A significant gap between the pooled and the separate score confirms the existence of structural or environmental differences between the clusters and the entire population, thus supporting the relevance of classifying the DMU into sub-groups (Rao *et al.*, 2003).

Coming to the variables used in clustering the DMUs, the levels of inputs and outputs are used in several studies aiming to address the scale heterogeneity occurring among units of different size (Amirteimoori and Kordrostami, 2013; Hajiagha *et al.*, 2016; Li *et al.*, 2016). The exclusive use of inputs and outputs allows for determining the extent of heterogeneity, even if it does not provide any information about what constitutes heterogeneity (Zarrin *et al.*, 2022).

In our study, we aim to address the heterogeneity related to some structural characteristics of the DMUs but also to exogenous factors related to the environment where the DMUs operate. The use of environmental variables, when needed, is hindered by the difficulty to identify and measure those variables (Dyson *et al.*, 2001). Moreover, since the number of environmental variables that are worth considering can be really large, the inclusion of too many variables into a DEA model would determine a reduction of its discriminatory capacity (Dyson *et al.*, 2001; Samoilenko and Osei-Bryson, 2010).

For what concerns the problems related to the identification and measurement of environmental variables, the ever-increasing development of Big Data sources may open new possibilities to retrieve both structured and unstructured data from internal and external sources.

Big Data is widely used in cluster analysis. However, despite significant interest in many research fields, its application in the retail sector, as well as in combination with DEA, remains limited [1]. Kulkarni *et al.* (2022) aim to deploy a Big Data model in the retail industry and use DEA to analyse the level of efficiency of different kinds of variables. In cluster analysis applications, Big Data is used, among others, to support predictions concerning the best store location (Andriyanov *et al.*, 2022; Carpio-Pinedo and Gutierrez, 2020; Robinson and Caradima, 2023) or store layout (Liao and Tasi, 2019), or it used to improve efficiency of e-commerce (Zatonatska *et al.*, 2022) or to assess the touristic attractivity of luxury store buildings (Pantano and Dennis, 2019): To the best of our knowledge this is the first study adopting a Big Data cluster analysis in combination DEA in the retail industry domain.

In particular, HMD and GIS are combined to provide a representation of external exogenous variables characterizing the context in which the DMUs operate that are worth considering for a reliable assessment of productivity.

The peculiarities of the Big Data sources employed also contribute to the originality of the approach followed since the combination of HMD and GIS allows considering not only the

characteristics of local residents in a specific area but also those of people temporarily passing through, such as tourists.

The case study used to test the data-driven clustering and DEA concerned a large kitchen furniture producer headquartered in Italy, whose managers needed to compare the productivity of hundreds of retailers distributed all over the country. Moreover, in Italy, as in other countries, remarkable differences can exist between northern and southern regions as well as between big cities and small towns. Such differences may be of demographic, social, economic and even cultural nature and may impact the productivity of different retailers.

Embedded with the use of Big Data is the necessity to reduce the dimensionality to improve the discriminatory capacity of the clustering technique (Allaoui *et al.*, 2020; Boutsidis *et al.*, 2014; Carlo *et al.*, 2019). PCA and MCA are well-established methods to reduce dimensionality (Tasoulis *et al.*, 2020), which we used in tandem with cluster analysis, aiming to both preserve the structure of Big Data and reduce the loss of information.

The DEA was then used within the clusters to compare retailers' productivity and support decision-making—for example, regarding the amount of money that the producer should invest in each retailer to strengthen commercial relations and improve company sales, as well as productivity and profitability overall.

Finally, the Big Data + DEA method was compared with a traditional DEA approach under a meta-frontier approach to facilitate a discussion about the results, accuracy and reliability of the proposed method.

In the next section, we describe the basics of DEA.

3. Data envelopment analysis

DEA is a non-parametric, data-driven benchmarking technique based on linear programming (Charnes *et al.*, 1978) that enables the comparison of productivity performance within a homogeneous group of DMUs. DMUs may correspond to stores, production centres, or any organizational units with a degree of decision-making autonomy (Zhu, 2022).

DEA provides a relative measure of productivity of every DMU, articulated as an output/input ratio:

$$\frac{Y}{X}$$

where Y and X represent the outputs produced and the related inputs, respectively.

The use of DEA for the performance measurement of a group of DMUs has several advantages. DEA allows us to sum up a multiplicity of outputs and inputs, even of different nature (e.g. internal or external and controllable or exogenous), while providing detailed information on the performance of every single factor (Gandhi and Shankar, 2014).

The multiple outputs and inputs are summarized as unique Y and X variables, respectively, using a weighted sum. The weights are estimated for each DMU through an optimization process aimed at maximizing the DMUs' productivity ratio. In formal terms, the linear programming problem to be solved for each DMU can be expressed as follows:

$$\text{Maximize } \frac{\sum_{r=1}^{s} u_r y_{rk}}{\sum_{i=1}^{m} v_i x_{ik}}$$

subject to

$$\frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} \leq 1 \, j = 1, \ldots, n$$

$$u_r, v_i > 0 \, \forall \, r = 1, \ldots, s; \; i = 1, \ldots, m$$

where.

n = number of DMUs

m = number of inputs

s = number of outputs

x_{ik} = quantity of input i consumed by DMU k

y_{rk} = quantity of output r produced by DMU k

v_i = weight of input i (unknown, to be determined)

u_r = weight of output r (unknown, to be determined)

The relative productivity score (ρ) of each DMU is obtained by comparing its Y/X ratio with the highest ratio within the analysed set of DMUs, which is considered a benchmark. In more formal terms, the productivity score of the *j-th DMU* can be calculated as follows:

$$\rho_j = \frac{\left(\frac{Y}{X}\right)_j}{\left(\frac{Y}{X}\right)_{max}}$$

Once the maximum output/input ratio has been established, it is possible to construct a "best practice frontier"—that is, a set of possible combinations of outputs and inputs according to the proportions of the most efficient DMUs.

A maximum productivity score of 1 (or 100%) is assigned to the DMUs located on the frontier. The score decreases in inverse proportion to the distance of the DMUs from the frontier (see Figure 1).

Suppose that the line passing through points A and B is the best practice frontier, where the angle corresponds to the output/input ratio. All DMUs located on this line are equally efficient and have the same optimal output/input ratio despite differences in the x and y variables. Conversely, DMU B′ exhibits a certain degree of inefficiency, being somewhat far from the frontier. More specifically, B′ has a lower output/input ratio, either compared to DMU B (the same output but a higher consumption of inputs) or to DMU A (a lower output with the same consumption of input).

The inefficiency of B′ can be derived differently depending on whether the DEA model is "input" or "output" oriented.

In an output-oriented approach, the assumption is that DMU B′ should tend to increase its efficiency by moving vertically along the direction of the segment $\overline{B'A}$ towards the best practice frontier. Therefore, the productivity score is obtained by comparing the length of the segments $\overline{IB'}$ and $\overline{IA}$:

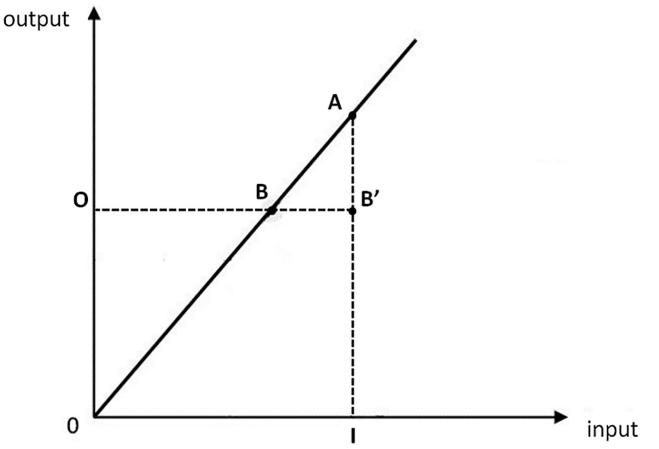


Figure 1.
Measurement of efficiency scores

Source(s): Authors' own creation

$$\rho_{B'} = \frac{\overline{IB'}}{\overline{IA}}$$

According to an input orientation, DMU B′ is expected to increase its productivity by reducing consumption and thus moving horizontally towards point B. Consequently, its productivity score is equal to the distance $\overline{OB}$ divided by the distance OB':

$$\rho_{B'} = \frac{\overline{OB}}{\overline{OB'}}$$

The projection (so called *radial*) of DMU B′ on the frontier represents the target unit to be achieved for it to become efficient. In the case of output-oriented approach, the (radial) target value for output is equal to $\theta * y_{B'}$, (holding constant the input $x_{B'}$), where θ is the reciprocal of the productivity score of DMU B'. Instead, in the case of input-oriented approach, the target value of input is equal to $\rho_{B'} * x_{B'}$, (holding constant the output $y_{B'}$).

Decision-makers can opt for an output-oriented rather than input-oriented DEA model according to the variables that they can influence to a greater extent. For example, if a fixed quantity of input is assigned to a DMU, decision-makers should aim to maximize the output, thus opting for an output orientation.

Regardless of the model's orientation, when the productivity score is lower than 1, the factors causing inefficiency need to be examined by referring back to the basic DEA model used to calculate the productivity ratio.

3.1 Constant return to scale (CRS) model vs variable return to scale (VRS) model

Before using DEA, one needs to decide whether a CRS or VRS model is more suitable. CRS models assume constant returns to scale in inputs and outputs, while VRS models are more coherent in the case of variable returns to scale. Furthermore, CRS and VRS can be developed into a variety of alternative models.

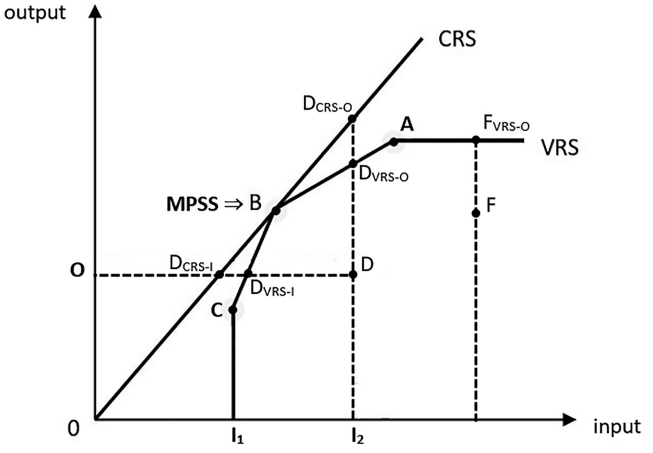
A comparison of these two types of models can be useful in revealing the causes (or sources) of DMU inefficiency.

Figure 2 shows the CRS and VRS frontiers. For all DMUs on the CRS frontier, it is assumed that an increase in input produces a proportional increase in output—that is, the angle of the line passing through the DMUs with the same efficiency is constant. Conversely, the VRS frontier assumes variable returns to scale, with different inclinations for different input levels. For example, if we have a DMU in the segment $\overline{CB}$, an increase in input will produce an increasing return to scale (IRS; in other words, the increase in output is proportionally higher than that of input). Contrariwise, for a DMU in the segment $\overline{BA}$, an increase in input will produce a proportionally lower increase in output (i.e. decreasing return to scale [DRS]). From a decision-maker perspective, it is worth investing (i.e. increasing input volumes) in DMUs that allow for IRS; in the case of DRS, diseconomies of scale must be reduced or solved to improve productivity.

Another crucial concept within the context of the VRS frontier is "slack." As depicted in Figure 2, DMU F is not positioned on the frontier. To achieve efficiency, it must first move to point $F_{VRS\text{-}O}$. When situated at this point, DMU F should have an efficiency score of 100% since it's now on the VRS frontier. However, DMU A, which is also on the frontier, produces the same output quantity with less input than DMU $F_{VRS\text{-}O}$, making it unable to achieve 100% efficiency. To attain a 100% efficiency score, DMU $F_{VRS\text{-}O}$ must move even further to point A. This additional improvement required for a DMU to reach efficiency is referred to as "slack." In fact, every DMU located along sections of the frontier that run parallel to either the x or *y* axes needs adjustments for slacks.

Slacks represent the potential improvements in input and output quantities for the inefficient units when compared with their "ultimate" benchmarks among efficient peers. In other words, they relate to the additional increases in output or reduction in input that can be achieved beyond what's indicated by the radial projection of inefficient units onto the frontier. VRS models are designed to account for these slacks.

The intersection between the CRS and VRS frontiers constitutes the most productive scale size (MPSS), which refers to the optimal size of a DMU with maximum efficiency (no diseconomies) and with all economies of scale being exploited.



Source(s): Authors' own creation

Figure 2.
The constant return to scale and variable return to scale frontier

The gap observed between the CRS and VRS frontiers entails a problem of scale. It is possible to measure the scale efficiency score (P_{scale}) as follows:

$$\rho_{scale} = \frac{\rho_{CRS}}{\rho_{VRS}}$$

where P_{CRS} and P_{VRS} are the CRS score and the VRS score, respectively. When P_{scale} is lower than 1, a DMU is not operating at optimal scale: if the DMU is located above the MPSS, it is using too many inputs and the scale is too large (see DMU A in Figure 2); when a DMU is located below the MPSS (see DMU C in Figure 2), the scale is too small.

The DMUs that are located neither on the CRS nor the VRS frontier may have simultaneous scale and management problems. See, for example, DMU D in Figure 2. Decision-makers should introduce initiatives to reduce managerial inefficiencies (i.e. increase output with the same input or reduce input for the same output) in order to move towards a VRS efficiency state, which corresponds to $D_{VRS\text{-}I}$ or $D_{VRS\text{-}O}$ on the VRS frontier.

Furthermore, DMU D should make scale adjustments to eliminate the scale problems and become CRS efficient, thus moving from point $D_{VRS\text{-}I}$ (or $D_{VRS\text{-}O}$) to point $D_{CRS\text{-}I}$ (or $D_{CRS\text{-}O}$) on the CRS frontier.

Therefore, a DMU can be inefficient under the CRS and VRS assumptions. CRS inefficiency is called *total inefficiency* and can be divided into VRS inefficiency (i.e. pure inefficiency) and scale inefficiency. These concepts can be expressed graphically using the three ratios, bounded by zero and one, according to the model's orientation (see Table 3):

4. Empirical application

4.1 Data

The tandem + DEA method was used in the context of an Italian kitchen furniture producer (the company hereafter) that was the market leader in terms of sales and turnover volumes.

As part of its strategy, the company was promoting a large-scale opening of new stores all over the country. In our analysis, all stores opened within the last 12 months (by June 2022) were considered, amounting to 541 stores in total. It is worth noting that although the stores were all owned by private retailers, the company contributed to their management through a commercial affiliation formula. More specifically, the company covered all costs related to store setup, while retailers committed to selling exclusively the kitchen furniture produced by the company, which is offered under two brands (Brand 1 and Brand 2) that are positioned in different market segments.

In recent years, the company has invested a remarkable amount of money in supporting the opening of new stores; consequently, controlling store productivity over time was essential in evaluating the company's investment returns.

The company's managers used to measure store productivity by considering two input variables and two output variables. The input variables were the average number of kitchen

Table 3. Ratio efficiency measures

	Total efficiency of D under CRS	Pure efficiency of D under VRS	Scale efficiency of D
Input-oriented DEA model	$\rho_{CRS-I} = \frac{\overline{OD_{CRS-I}}}{\overline{OD}}$	$\rho_{VRS-I} = \frac{\overline{OD_{VRS-I}}}{\overline{OD}}$	$\rho_{scale-I} = \frac{\overline{OD_{CRS-I}}}{\overline{OD_{VRS-I}}}$
Output-oriented DEA model	$\rho_{CRS-O} = \frac{\overline{ID}}{\overline{ID_{CRS-O}}}$	$\rho_{VRS-O} = \frac{\overline{ID}}{\overline{ID_{VRS-O}}}$	$\rho_{scale-O} = \frac{\overline{ID_{CRS-O}}}{\overline{ID_{VRS-O}}}$

models presented in the store and the total costs of the store setup, while sales expressed both in terms of volume and turnover were included as outputs.

A general threshold of minimum expected productivity was defined as a benchmark for evaluating each store. However, the results of the initial analysis revealed remarkable differences in store performance that could not be simply attributed to management inefficiencies or inadequate scale. Consequently, the company managers realized the importance of refining the analysis by grouping the stores according to similar internal characteristics and operating environments before making judgements on the performance of individual stores.

To this end, three variables representing the stores' structural characteristics were considered, together with 13 variables representing the socio-demographic characteristics of a store's catchment area.

The catchment area for each store was defined by using a 30-min drive-time isochrone, which the management considered to represent the maximum distance that a potential customer (a prospect hereafter) would be willing to cover to reach a kitchen store.

The information on the socio-economic and demographic attributes of each catchment area was collected using spatial Big Data. More specifically, data sourced via GIS were combined with the HMD provided by a leading telecommunications company. The integration of GIS data and HMD provided accurate data about the actual population living in a determined area, which could be significantly different from the population of formal residents. A list of all the variables used for store clustering is provided in Table 4.

4.2 Factorial methods and clustering

An MCA was performed for the six categorical variables (nominal and binary in Table 4) to extract the factors that effectively summarized the data (see Table 5).

Variable	Description	Data source
Store type	Nominal: Flagship store Brand 1/2/1 + 2; Exclusive reseller Brand 1/2/1 + 2; Non-exclusive reseller	Internal
Store's renewal level	Nominal: new (open in the last 5 years); renewed (renewed in the last 5 years); old (not open or renewed in the last 5 years)	Internal
Recent update of kitchen models presented in the store	Binary	Internal
Proximity to shopping centres	Binary	GIS
Proximity to department stores	Binary	GIS
Proximity to parking lots	Binary	GIS
Inhabitants	Discrete	HMD
Catchment area (sq. km)	Continuous	GIS
Population density	Continuous	HMD + GIS
Percentage of target population for Brand 1	Continuous	HMD
Percentage of target population for Brand 2	Continuous	HMD
Purchasing power per inhabitant	Continuous	HMD + GIS
Furniture expenses per inhabitant	Continuous	HMD + GIS
Kitchen furniture expenses per inhabitant	Continuous	HMD + GIS
Number of competitor stores	Discrete	GIS
Number of large furniture stores	Discrete	GIS

Table 4. Variables used for stores profiling stores

The first two factors explained 60.1% of the adjusted inertia (Greenacre, 1993)—of the information contained in the data. Each factor was interpreted by examining the contribution of each category (of the variables analysed) to the two factors by considering the corresponding squared cosines to avoid misinterpretations (if, for a given category, the cosines are low in relation to the factor of interest, then an interpretation is hazardous).

As shown in Table 6, the first factor was correlated with a store's renewal level, which was, in turn, associated with a recent update of the kitchen models presented in the store and the presence of the company's brands on the store's signboard. The second factor was correlated to a store's proximity to shopping centres and parking lots.

A PCA was performed for the 11 quantitative variables, with the same objective as for the MCA. According to the "scree test" method (Cattell, 1966), three factors were extracted, which represented 87.3% of the cumulative variability (see Table 7).

The contributions of each variable in building the factors and the squared cosines were used to interpret the results. The first factor was linked to the intensity of competition within the catchment area, which was, in turn, correlated with the general purchasing power of its inhabitants and per capita spending on furniture. The second factor can be seen as a measure of the presence of the target population for Brand 2 within the catchment area. The third factor was correlated with the presence of the target population for Brand 1 (see Table 8).

Table 5. Multiple correspondence analysis: eigenvalues and adjusted inertia

Factor	Eigenvalues	Adjusted inertia (%)	Adjusted inertia (% cum.)
F1	0.275	54.8	54.8
F2	0.200	5.3	60.1
F3	0.190	2.6	62.7
F4	0.178	0.6	63.2
F5	0.171	0.1	63.3
F6	0.167	0.0	63.3

Table 6. Multiple correspondence analysis: principal coordinates, contributions and squared cosines

Variable–category	Principal coordinates		Contributions		Squared cosines	
	F1	F2	F1	F2	F1	F2
Store's sign–Flagship store Brand 1	−0.624	−0.173	0.004	0.000	0.007	0.001
Store's sign–Flagship store Brand 2	−0.914	2.498	0.022	0.221	0.037	0.277
Store's sign–Flagship store Brand 1 + 2	−0.963	−0.437	0.087	0.025	0.170	0.035
Store's sign–Non-exclusive retailer	−0.707	−0.548	0.060	0.049	0.123	0.074
Store's sign–Exclusive retailer Brand 1	1.284	−0.205	*0.131*	0.005	*0.249*	0.006
Store's sign–Exclusive retailer Brand 2	0.895	−0.051	*0.115*	0.001	*0.248*	0.001
Store's sign–Exclusive retailer Brand 1 + 2	−0.186	0.515	0.005	0.048	0.010	0.074
Store–new	1.141	−0.118	*0.283*	0.004	*0.727*	0.008
Store–renovated	0.643	0.122	0.004	0.000	0.006	0.000
Store–old	−0.668	0.065	0.170	0.002	0.749	0.007
Showroom–new	1.048	−0.454	0.084	0.022	0.158	0.030
Showroom–old	−0.151	0.065	0.012	0.003	0.158	0.030
Proximity to shopping centres–no	−0.028	−0.122	0.000	0.012	0.019	0.366
Proximity to shopping centres–yes	0.683	3.009	0.011	*0.292*	0.019	*0.366*
Proximity to department stores–no	0.001	−0.009	0.000	0.000	0.001	0.022
Proximity to department stores–yes	−0.389	2.444	0.000	0.018	0.001	0.022
Proximity to parking lots–no	−0.057	−0.249	0.002	0.044	0.019	0.358
Proximity to parking lots–yes	0.331	1.436	0.010	*0.254*	0.019	*0.358*

The MCA and PCA were used as part of a pre-processing step for classification purposes: the coordinate values of the stores related to the factors derived from the MCA and PCA were used as input for clustering algorithms. AHC was performed (Euclidean distance, Ward's method) to determine the optimal number of clusters (Everitt, 1993). The inspection of the dendrogram produced by the AHC suggested the generation of three clusters. A k-means algorithm was then used to group the stores into three clusters. Subsequently, the test-values of the variables (arranged in descending order) were used to identify the most discriminant variables in the characterization of clusters and to create the corresponding profiles (Lebart, 2000; Villanueva *et al.*, 2013).

As shown in Table 9, Cluster 1 was characterized by stores being located in populous catchment areas with strong competitive pressure (the number of inhabitants and the number of competitor stores were remarkably higher than the average). In Cluster 2, the stores operated in catchment areas that were richer than average (high purchasing power per inhabitant and high per capita expenses for kitchens and furniture). Cluster 3 included the stores located in catchment areas where the presence of the target population for Brand 2 was higher than the average.

4.3 Results and discussion

First, DEA was performed on the entire set of DMUs using the DeaFrontier software. The output-oriented Banker, Charnes, Cooper's model was chosen in coherence with the company's objective to maximize sales by emphasizing the "pure inefficiency" attributable to

Factor	Eigenvalues	Variability (%)	Variability (% cum.)
F1	5.457	54.568	54.568
F2	2.370	23.697	78.265
F3	0.905	9.053	87.318
F4	0.607	6.067	93.385
F5	0.286	2.858	96.243
F6	0.184	1.837	98.080
F7	0.108	1.078	99.158
F8	0.055	0.547	99.705
F9	0.019	0.192	99.897
F10	0.010	0.103	100.000

Table 7.
Principal component analysis: eigenvalues and variability

	Contributions %			Squared cosines		
Variable	F1	F2	F3	F1	F2	F3
Inhabitants	14.083	7.686	2.995	0.769	0.182	0.027
Catchment area (sq. km)	9.617	0.432	0.083	0.525	0.010	0.001
Population density	10.652	9.153	3.353	0.581	0.217	0.030
% of target population for Brand 1	3.188	1.421	*83.722*	0.174	0.034	*0.758*
% of target population for Brand 2	1.789	*31.972*	2.581	0.098	*0.758*	0.023
Purchasing power per inhabitant	9.529	18.932	0.204	0.520	0.449	0.002
Furnishing expenses per inhabitant	9.590	18.177	0.059	0.523	0.431	0.001
Kitchen furniture expenses per inhabitant	12.768	2.412	3.119	0.697	0.057	0.028
Number of competitor stores	13.649	8.432	3.376	0.745	0.200	0.031
Number of large furniture stores	*15.133*	1.383	0.508	*0.826*	0.033	0.005

Table 8.
Principal component analysis: contributions (%) and squared cosines

Variable	Cluster value	Mean value	Test value
Cluster 1			
Inhabitants	2,538,812	*591,289*	20.359
Number of competitor stores	78.1	*19.3*	20.185
Number of large furniture stores	10.2	*2.6*	18.270
Population density	1,985	*594*	18.201
Catchment area (sq. km)	1,352	*821*	13.504
Cluster 2			
Purchasing power per inhabitant	18,859	*16,329*	13.482
Furnishing expenses per inhabitant	325	*286*	12.454
Cluster 3			
% target population Brand 2	12.1%	*11.2%*	23.189

Table 9.
Cluster profiling: variables with test values >10

store management. The model used can be expressed as a linear programming problem in enveloped form as follows:

$$Maximize\ \phi_k + \varepsilon \sum_{r=1}^{s} s_r + \varepsilon \sum_{i=1}^{m} s_i$$

subject to

$$\phi_k y_{rk} - \sum_{j=1}^{n} \lambda_j y_{rj} + s_r = 0\ r = 1, \ldots, s$$

$$x_{ik} - \sum_{j=1}^{n} \lambda_j x_{ij} - s_i = 0\ i = 1, \ldots, m$$

$$\sum_{j=1}^{n} \lambda_j = 1$$

$$\lambda_j, s_r, s_i \geq 0\ \forall j = 1, \ldots, n;\ r = 1, \ldots, s;\ i = 1, \ldots, n$$

where.

x_{ij} = quantity of input i consumed by the *j-th* DMU

y_{rj} = quantity of output r produced by the *j-th* DMU

λ_j = weights of outputs and inputs of the *j-th* DMU

s_i = input slacks

s_r = output slacks

= non-Archimedean value (smaller than any positive real number and greater than 0)

The assumption of VRS, as expressed through the third constraint in the model, enables us to focus on the technical efficiency of DMUs, i.e. how they utilize available resources, without considering any scale inefficiency. The efficient targets for outputs and inputs (including slacks) are calculated as follows:

$$Outputs: \widehat{y}_{rk} = \phi y_{rk} + s_r\ r = 1, \ldots, s$$

$$Inputs: \widehat{x}_{ik} = x_{ik} - s_i \; i = 1, \ldots, m$$

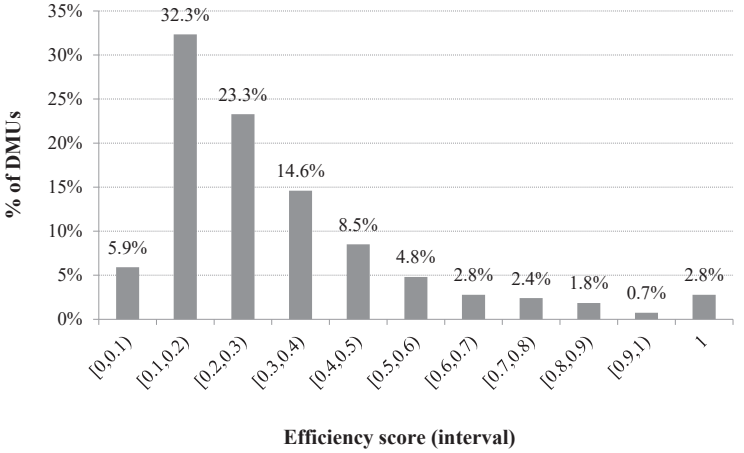
As shown in Figure 3, under a meta-frontier approach, where efficiency scores are calculated for the entire dataset, only 2.8% of the stores achieved a full efficiency score, while the remaining 97.2% were inefficient. The average efficiency score was 0.313, with 84.7% of the stores scoring below 0.5. If these initial results were reliable, then the retailers were highly inefficient, and a significant improvement would have been required to reach maximum efficiency (score = 1).

The second step involved conducting DEA on each of the three clusters separately. The local (or separated) efficiency scores were subsequently compared to the pooled scores. It was evident that the latter were lower than the former, as theoretically postulated. However, the difference between the local and the pooled score was significant in most of the cases, with over half of the DMUs showing a difference exceeding 20% (see Table 10).

As shown in Table 10, the average gap ratios of the three clusters ranged between 0.822 and 0.894. This means that the factors defining the cluster profile had a significant impact on the efficiency scores.

Figure 4 provides an overview of the distribution of all the DMUs in the three clusters concerning the comparison between the separated (local) and the pooled productivity scores.

In particular, the pooled scores arranged in ascending order are distributed along the line represented by the black dots, while the red dots represent the corresponding local scores.



Source(s): Authors' own creation

Figure 3.
Interval distribution of DMU meta-frontier (or pooled) efficiency scores

	Meta-frontier DEA	Local DEA		
		Cluster 1	Cluster 2	Cluster 3
Number of stores	541	75	229	237
Number of efficient stores	15	6	9	10
Percentage of efficient stores	2.8%	8.0%	3.9%	4.2%
Average meta-frontier efficiency score	0.313	0.386	0.308	0.295
Average local efficiency score		0.461	0.346	0.353
Efficiency gap ratio		0.845	0.894	0.822

Table 10.
Comparison of meta-frontier and local efficiency scores

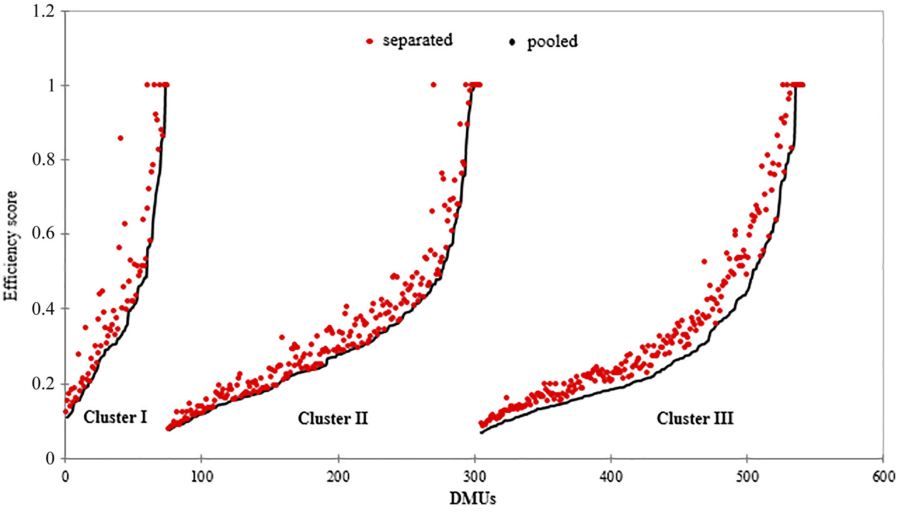


Figure 4.
Distribution of separated and pooled productivity scores in the three clusters (Nr-total DMUs = 541)

Source(s): Authors' own creation

The cluster formed by the red dots is generally situated above the black line, indicating that the productivity assessment generally increases if calculated on the group-frontier rather than on the meta-frontier. This is especially evident in cluster 3, where the red dots aligning with the black line (local score = pooled score) are limited to very few cases.

In all three clusters, there are some DMUs with a local efficiency score that significantly differs from the corresponding pooled score. For these DMUs, it's clear how the choice between local or separate scoring dramatically affects the expression of efficiency assessment.

Given that the aim of clustering is to maximize the homogeneity of the stores within the same group, the efficiency scores obtained using this approach can be considered more reliable, which can help setting more attainable targets, particularly for less efficient retailers. As previously discussed, local productivity scores are higher than the meta-frontier scores. Therefore, when assessing productivity using local scores, the distance of a DMU from the local efficiency frontier is reduced compared to when using meta-frontier scores. In other words, transitioning from pooled productivity scores to local scores results in a relative reduction in the effort required for a non-efficient DMU to reach the maximum level of efficiency. An example is provided in Figure 5, where the vertical lines represent the percentage reduction in the target set for each DMU when switching from pooled productivity scores to local productivity scores for the "revenues" variable (the target values are provided by the DEAFrontier software).

As Figure 5 illustrates, the reduction in local targets compared to pooled targets is substantial for a significant number of DMUs, especially for those in clusters 1 and 3. In this sense, the use of local DEA scores enables a more realistic and attainable target setting. A similar distribution is observed for the "sales quantity" output variable.

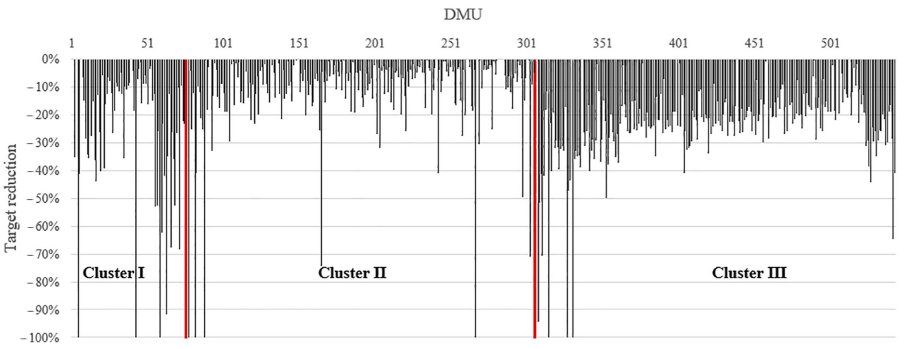
Table 11 summarizes the distribution of percentage reduction intervals for local targets compared to pooled targets. In Cluster 1, for 13% of retailers, the targets determined using local scores are more than 50% lower than the targets set based on pooled scores. The extent of reduction of local targets is influenced by the structural and environmental variables that characterize the clusters. In Cluster 1, the presence of a high number of competitors and large

furniture stores in the same area hinders the potential for improvement for the DMUs in that cluster. Cluster 2 is characterized by individual purchasing power and furnishing expenses per inhabitant slightly higher than the mean. As a result, the magnitude of target reduction is lower than in Cluster 1. The percentage reduction of local targets is less than 20% for approximately 88% of the DMUs. Finally, in Cluster 3, the reduction of local targets ranges from −20% to −30% for 40% of DMUs, due to the higher presence of individuals targeted for Brand 2.

The comparison of two examples of DMUs, named X and Y, provides further details about the usefulness of cluster and DEA in support of target settings (see Table 12).

DMU X was considered rather inefficient if evaluated through pooled efficiency score (meta-frontier DEA), while it scored 100% of local efficiency. The target set on benchmarking the pooled score on the meta-frontier would require the DMU to achieve an improvement of 77% on its actual results, while the target set on the local frontier would not require any improvement, since the DMU is already on the maximum achievable target of efficiency.

As for DMU Y, the pooled efficiency score of 0.33 would necessitate a substantial increase of 203% in revenues and sales volume, which might be perceived as unattainable. Conversely, DMU Y is much closer to the local frontier (local score = 0.86), reducing the magnitude of effort needed to reach the maximum efficiency frontier. Consequently, more realistically achievable targets can be more effective in motivating managers.



Source(s): Authors' own creation

Figure 5.
Percentage reduction of targets calculated through "local" versus "pooled" productivity scores (variable: revenues)

% variation	Cluster 1 (% of DMUs)	Cluster 2 (% of DMUs)	Cluster 3 (% of DMUs)	% of total DMUs
0%	8.0%	3.9%	4.6%	4.8%
[0%, −10%)	26.7%	58.5%	7.6%	31.8%
[−10%, −20%)	22.7%	26.2%	28.3%	26.6%
[−20%, −30%)	13.3%	6.1%	40.9%	22.4%
[−30%, −40%)	10.7%	1.3%	11.4%	7.0%
[−40%, −50%)	5.3%	1.3%	4.2%	3.1%
[−50%, −60%)	4.0%	0.0%	0.4%	0.7%
[−60%, −70%)	4.0%	0.0%	0.4%	0.7%
[−70%, −80%)	0.0%	0.9%	0.4%	0.6%
[−80%, −90%)	0.0%	0.0%	0.0%	0.0%
[−90%, −100%)	1.3%	0.0%	0.4%	0.4%
−100%	4.0%	1.7%	1.3%	1.8%

Table 11.
Interval distribution of % reduction of local versus pooled targets (variable: revenues)

Table 12.
Local versus pooled targets. Comparison of DMU X and Y

DMU name	X	Y
Cluster	1	1
Productivity score (pooled)	0.56	0.33
Productivity score (local)	1.00	0.86
Actual output 1 (revenues €)	160,014	137,290
Actual output 2 (sales in quantity)	37	28
Target output 1 (pooled)	283,341	415,454
Target output 2 (pooled)	66	85
Improvement (%) Output 1 (pooled)	77%	203%
Improvement (%) Output 2 (pooled)	77%	203%
Target Output 1 (local)	160,014	160,014
Target Output 2 (local)	37	37

5. Conclusions

Research on efficiency and productivity in the retail industry has generated significant interest among scholars and practitioners. In this field, several studies have attempted to improve measurement accuracy by means of complex techniques derived from statistics and other quantitative disciplines. DEA is extensively used for this purpose due to its numerous advantages, including its flexibility across a wide range of applications. It enables the analysis of multiple aspects of productivity by summarizing diverse input and output variables of varying nature. Furthermore, it provides relative measures of productivity or efficiency that prove effective in supporting decision-making. However, the heterogeneity within the set of DMUs analysed is a major concern in DEA applications, which can impact result accuracy. This issue is mitigated by grouping DMUs into homogeneous sub-groups using cluster analysis. In the literature, cluster analysis is predominantly employed to address scale heterogeneity among DMUs, whereas studies focused on the heterogeneity resulting from external exogenous variables are relatively limited. This limitation arises, in part, from challenges in identifying suitable measures and the potential loss of discriminatory power in the model when incorporating too many variables (dimensionality).

The growing use of Big Data may expand the possibilities of introducing new measures related to both internal and external variables. These measures can be applied to enhance a cluster analysis combined with DEA with the goal of reducing the heterogeneity within a set of DMUs driven by external environmental characteristics as well as internal structural characteristics.

This article contributes to this direction. We described a measurement technique based on a case study of a leading kitchen furniture manufacturer. The management team needed a method to assess productivity at the store level within a commercial network consisting of over 500 retailers for the purposes of performance-measurement and target-setting.

The productivity index was measured using an output-oriented DEA model. Big Data was used in the steps preceding DEA, where a combination of GIS and HMD enabled a tandem analysis used to divide the retailers into clusters that represent some of the retailers' structural characteristics, and their competitive and demographic and demographic landscape.

In particular, the tandem analysis consists of a PCA and an MCA used to address the dimensionality stemming from the numerous available variables. These analyses reduce the variables into significant factors, enabling us to uncover the essential phenomena behind the raw data while minimizing information loss. The resulting factors are then applied in the clustering algorithm, which utilizes AHC and k-means.

The DEA was conducted within the clusters, following a meta-frontier approach in line with the aim of analysis. We also applied DEA to the entire set of DMUs, allowing us to

compare "local" and "pooled" efficiency scores, facilitating discussions about the accuracy and reliability of results and their practical use in decision-making.

This paper contributes to various research directions. The tandem analysis enables the inclusion of exogenous external variables in cluster + DEA applications while simultaneously reducing dimensionality, which is particularly relevant when dealing with Big Data.

The use of HMD represents a novel approach that allows us to profile a specific area based on the characteristics of the current population residing in that area, even for a limited period, while excluding individuals who are formally registered in the area but may reside elsewhere.

From a managerial perspective, the results reveal the significance of environmental variables in assessing the potential productivity of retailers, setting realistic improvement targets and identifying scale problems and inefficiencies that could be addressed through suitable initiatives. Company managers can use this information to determine the investment worthiness of a retailer or the expected maximum return on investment, considering the specific characteristics of the retailer's operating environment.

Although Big Data are largely adopted in clustering analysis, their use in the retail sector and cluster + DEA is still limited.

However, the proposed methodology is not without limitations, which also open up opportunities for future developments. The MCA/PCA factorial techniques are well established methods that have proven effective in a relevant number of cases, even if the representativeness of the structure of data in the clusters may not be guaranteed. Recently, more sophisticated clustering methods have been proposed in the literature that could potentially offer greater effectiveness. In this initial attempt, we prioritized the applicability of the method in a real-world context, making MCA/PCA a suitable choice due to their wide usage and ease of adoption. Nevertheless, the potential applicability of more advanced cluster techniques could be explored in future applications.

Furthermore, the methodology proposed is generalizable to all contexts dealing with the measurement and comparison of performance in commercial networks. However, users must be aware that the DEA methodology is based on best practices, consequently the obtained results are sensitive to the specific case and may vary depending on the size of the population under investigation and the analytical context. Thus, it may be necessary to adapt the methodology and the Big Data sources used when applying it to different industries. Future adoption of the tandem + DEA in various industries will be valuable for validating the method and assessing its generalizability.

Note

1. A Scopus search was conducted by using the research string: "Big Data AND Cluster Analysis", which yielded 1,875 documents, the majority of which are related to the research domains of computer science, engineering and mathematics. A second search was performed by using the string "Big Data AND Cluster Analysis AND Retail," and only 11 documents were selected, none of which using the DEA method. Finally, we searched "Big Data AND DEA AND retail" and only one document was retrieved (all the research were conducted on the 1st of October 2023).

References

Allaoui, M., Kherfi, M.L. and Cheriet, A. (2020), "Considerably improving clustering algorithms using UMAP dimensionality reduction technique: a comparative study", *International conference on image and signal processing*, Cham, Springer International Publishing, pp. 317-325.

Amirteimoori, A. and Kordrostami, S. (2013), "An alternative clustering approach: a DEA-based procedure", *Optimization*, Vol. 62 No. 2, pp. 227-240, doi: 10.1080/02331934.2011.585466.

Andriyanov, N., Dementiev, V., Tashlinsky, A. and Danilov, A. (2022), "Machine learning technologies for bakery management decisions", *2022 24th International Conference on Digital Signal Processing and its Applications (DSPA)*, pp. 1-6.

Banker, R.D., Charnes, A. and Cooper, W.W. (1984), "Some models for estimating technical and scale inefficiencies in data envelopment analysis", *Management Science*, Vol. 30 No. 9, pp. 1078-1092, doi: 10.1287/mnsc.30.9.1078.

Banker, R.D. and Morey, R.C. (1986), "The use of categorical variables in data envelopment analysis", *Management Science*, Vol. 12, pp. 1613-1627, doi: 10.1287/mnsc.32.12.1613.

Barros, C.P. (2005), "Efficiency in hypermarket retailing: a stochastic Frontier model", *The International Review of Retail, Distribution and Consumer Research*, Vol. 15 No. 2, pp. 171-189, doi: 10.1080/09593960500049381.

Bhat, S., Gijo, E.V. and Jnanesh, N.A. (2016), "Productivity and performance improvement in the medical records department of a hospital: an application of Lean Six Sigma", *International Journal of Productivity and Performance Management*, Vol. 65 No. 1, pp. 98-125, doi: 10.1108/ijppm-04-2014-0063.

Boutsidis, C., Zouzias, A., Mahoney, M.W. and Drineas, P. (2014), "Randomized dimensionality reduction for k-means clustering", *IEEE Transactions on Information Theory*, Vol. 61 No. 2, pp. 1045-1062, doi: 10.1109/tit.2014.2375327.

Carlo, C., Maurizio, V. and Zaccaria, G. (2019), "Hierarchical clustering and dimensionality reduction for big data", in *Smart Statistics for Smart Applications, Book of Short Paper SIS2019*, Pearson, pp. 173-180.

Carpio-Pinedo, J. and Gutiérrez, J. (2020), "Consumption and symbolic capital in the metropolitan space: integrating 'old' retail data sources with social big data", *Cities*, Vol. 106, 102859, doi: 10.1016/j.cities.2020.102859.

Cattell, R.B. (1966), "The scree test for the number of factors", *Multivariate Behavioral Research*, Vol. 1 No. 2, pp. 245-276, doi: 10.1207/s15327906mbr0102_10.

Charnes, A. and Cooper, W.W. (1980), "Auditing and accounting for program efficiency and management efficiency in not-for-profit entities", *Accounting, Organizations and Society*, Vol. 5 No. 1, pp. 87-107, doi: 10.1016/0361-3682(80)90025-2.

Charnes, A., Cooper, W.W. and Rhodes, E. (1978), "Measuring the efficiency of decision making units", *European Journal of Operational Research*, Vol. 2 No. 6, pp. 429-444, doi: 10.1016/0377-2217(78)90138-8.

Charnes, A., Cooper, W.W., Golany, B., Seiford, L. and Stutz, J. (1985), "Foundations of data envelopment analysis for Pareto-Koopmans efficient empirical production functions", *Journal of Econometrics*, Vol. 30 Nos 1-2, pp. 91-107, doi: 10.1016/0304-4076(85)90133-2.

Cinaroglu, S. (2020), "Integrated k-means clustering with data envelopment analysis of public hospital efficiency", *Health Care Management Science*, Vol. 23 No. 3, pp. 325-338, doi: 10.1007/s10729-019-09491-3.

Costa, M.A., Mineti, L.B., Mayrink, V.D. and Lopes, A.L.M. (2019), "Bayesian detection of clusters in efficiency score maps: an application to Brazilian energy regulation", *Applied Mathematical Modelling*, Vol. 68, pp. 66-81, doi: 10.1016/j.apm.2018.11.009.

Donthu, N. and Yoo, B. (1998), "Retail productivity assessment using DEA", *Journal of Retailing*, Vol. 74 No. 1, pp. 89-105, doi: 10.1016/s0022-4359(99)80089-x.

Dubelaar, C., Bhargava, M. and Ferrarin, D. (2002), "Measuring retail productivity: what really matters?", *Journal of Business Research*, Vol. 55 No. 5, pp. 417-426, doi: 10.1016/s0148-2963(00)00160-0.

Dyson, R.G., Allen, R., Camanho, A.S., Podinovski, V.V., Sarrico, C.S. and Shale, E.A. (2001), "Pitfalls and protocols in DEA", *European Journal of Operational Research*, Vol. 132 No. 2, pp. 245-259, doi: 10.1016/s0377-2217(00)00149-1.

Ebrahimnejad, A. and Amani, N. (2021), "Fuzzy data envelopment analysis in the presence of undesirable outputs with ideal points", *Complex and Intelligent Systems*, Vol. 7 No. 1, pp. 379-400, doi: 10.1007/s40747-020-00211-x.

Ebrahimnejad, A. and Lotfi, F.H. (2012), "Equivalence relationship between the general combined-oriented CCR model and the weighted minimax MOLP formulation", *Journal of King Saud University-Science*, Vol. 24 No. 1, pp. 47-54, doi: 10.1016/j.jksus.2010.08.007.

Ebrahimnejad, A. and Tavana, M. (2014), "An interactive MOLP method for identifying target units in output-oriented DEA models: the NATO enlargement problem", *Measurement*, Vol. 52, pp. 124-134, doi: 10.1016/j.measurement.2014.03.016.

Everitt, B.S. (1993), *Cluster Analysis*, Edward Arnold, London.

Farrell, M.J. (1957), "The measurement of productive efficiency", *Journal of the Royal Statistical Society Series A*, Vol. 120 No. 3, pp. 253-290, doi: 10.2307/2343100.

Foster, L., Haltiwanger, J. and Krizan, C.J. (2002), "The link between aggregate and micro productivity growth: evidence from retail trade", NBER Working Paper No. 9120.

Gandhi, A. and Shankar, R. (2014), "Efficiency measurement of Indian retailers using data envelopment analysis", *International Journal of Retail and Distribution Management*, Vol. 42 No. 6, pp. 500-520, doi: 10.1108/ijrdm-10-2012-0094.

Gandhi, A. and Shankar, R. (2016), "Strategic resource management model and data envelopment analysis for benchmarking of Indian retailers", *Benchmarking*, Vol. 23 No. 2, pp. 286-312, doi: 10.1108/bij-02-2014-0013.

Gauri, D.K. (2013), "Benchmarking retail productivity considering retail pricing and format strategy", *Journal of Retailing*, Vol. 89 No. 1, pp. 1-14, doi: 10.1016/j.jretai.2012.09.001.

Gong, Y., Liu, J. and Zhu, J. (2019), "When to increase firms' sustainable operations for efficiency? A data envelopment analysis in the retailing industry", *European Journal of Operational Research*, Vol. 277 No. 3, pp. 1010-1026, doi: 10.1016/j.ejor.2019.03.019.

Greenacre, M.J. (1993), "Biplots in correspondence analysis", *Journal of Applied Statistics*, Vol. 20 No. 2, pp. 251-269, doi: 10.1080/02664769300000021.

Griffith, R. and Harmgart, H. (2005), "Retail productivity", *International Review of Retail Distribution and Consumer Research*, Vol. 15 No. 3, pp. 281-290, doi: 10.1080/09593960500119481.

Günter, A. and Gopp, E. (2021), "Overview and classification of approaches to productivity measurement", *International Journal of Productivity and Performance Management*, Vol. 71 No. 4, pp. 1221-1229, doi: 10.1108/ijppm-05-2019-0241.

Gupta, A. and Mittal, S. (2010), "Measuring retail productivity of food and grocery retail outlets using the DEA technique", *Journal of Strategic Marketing*, Vol. 18 No. 4, pp. 277-289, doi: 10.1080/09652540903537055.

Haas, D.A. and Murphy, F.H. (2003), "Compensating for non-homogeneity in decision-making units in data envelopment analysis", *European Journal of Operational Research*, Vol. 144 No. 3, pp. 530-544, doi: 10.1016/s0377-2217(02)00139-x.

Hajiagha, S.H.R., Hashemi, S.S. and Mahdiraji, H.A. (2016), "Fuzzy C-means based data envelopment analysis for mitigating the impact of units' heterogeneity", *Kybernetes*, Vol. 45 No. 3, pp. 536-551, doi: 10.1108/k-07-2015-0176.

Haskel, J. and Khawaja, N. (2003), "Productivity in UK retailing: evidence from micro data", CERIBA Working Paper, available at: http://www.qmul.ac.uk/~ugte153/CERIBA/publications/services.pdf

Higón, D.A., Bozkurt, Ö., Clegg, J., Grugulis, I., Salis, S., Vasilakos, N. and Williams, A.M. (2010), "The determinants of retail productivity: a critical review of the evidence", *International Journal of Management Reviews*, Vol. 12 No. 2, pp. 201-217, doi: 10.1111/j.1468-2370.2009.00258.x.

Huang, H., Li, S. and Yu, Y. (2019), "Evaluation of the allocation performance in a fashion retail chain using data envelopment analysis", *The Journal of The Textile Institute*, Vol. 110 No. 6, pp. 901-910, doi: 10.1080/00405000.2018.1532376.

Ingene, C.A. (1984), "Productivity and functional shifting in spatial retailing: private and social perspectives", *Journal of Retailing*, Vol. 60, pp. 15-36.

Islam, S. and Syed Shazali, S.T. (2011), "Determinants of manufacturing productivity: pilot study on labor-intensive industries", *International Journal of Productivity and Performance Management*, Vol. 60 No. 6, pp. 567-582, doi: 10.1108/17410401111150751.

Jiang, H., Wu, J., Chu, J. and Liu, H. (2020), "Better resource utilization: a new DEA bi-objective resource reallocation approach considering environmental efficiency improvement", *Computers and Industrial Engineering*, Vol. 144, 106504, doi: 10.1016/j.cie.2020.106504.

Kamakura, W.A., Lenartowicz, T. and Ratchfrord, B.T. (1996), "Productivity assessment of multiple retail outlets", *Journal of Retailing*, Vol. 72 No. 4, pp. 333-356, doi: 10.1016/s0022-4359(96)90018-4.

Käpylä, J., Jääskeläinen, A. and Lönnqvist, A. (2010), "Identifying future challenges for productivity research: evidence from Finland", *International Journal of Productivity and Performance Management*, Vol. 59 No. 7, pp. 607-623, doi: 10.1108/17410401011075620.

Keh, H.T. and Chu, S. (2003), "Retail productivity and scale economies at the firm level: a DEA approach", *Omega*, Vol. 31 No. 2, pp. 75-82, doi: 10.1016/s0305-0483(02)00097-x.

Ko, K., Chang, M., Bae, E.S. and Kim, D. (2017), "Efficiency analysis of retail chain stores in Korea", *Sustainability (Switzerland)*, Vol. 9 No. 9, pp. 1-14, doi: 10.3390/su9091629.

Kulkarni, P.M., Gokhale, P. and Dandannavar, P.S. (2022), "Big data challenges in retail sector: perspective from data envelopment analysis", *In International Conference on Big Data Innovation for Sustainable Cognitive Computing*, pp. 89-97.

Kumar, V., Anand, A. and Song, H. (2017), "Future of retailer profitability: an organizing framework", *Journal of Retailing*, Vol. 93 No. 1, pp. 96-119, doi: 10.1016/j.jretai.2016.11.003.

Lebart, L. (2000), "Contiguity analysis and classification", *Data Analysis: Scientific Modeling and Practical Application*, pp. 233-243.

Li, Y., Hou, S.H. and Yao, L.M. (2016), "Profitability assessment using data envelopment with cluster analysis: a case for different types of gas stations", *Chemical Engineering Transactions*, Vol. 51, pp. 727-732.

Liao, S.H. and Tasi, Y.S. (2019), "Big data analysis on the business process and management for the store layout and bundling sales", *Business Process Management Journal*, Vol. 25 No. 7, pp. 1783-1801, doi: 10.1108/bpmj-01-2018-0027.

Lozano-Vivas, A., Pastor, J.T. and Pastor, J.M. (2002), "An efficiency comparison of European banking systems operating under different environmental conditions", *Journal of Productivity Analysis*, Vol. 18 No. 1, pp. 59-77, doi: 10.1023/a:1015704510270.

Majdi, M., Ebrahimnejad, A. and Azizi, A. (2023), "Common-weights fuzzy DEA model in the presence of undesirable outputs with ideal and anti-ideal points: development and prospects", *Complex & Intelligent Systems*, pp. 1-18, doi: 10.1007/s40747-023-01030-6.

Mishra, A. and Ansari, J. (2013), "A conceptual model for retail productivity", *International Journal of Retail and Distribution Management*, Vol. 41 No. 7, pp. 348-379, doi: 10.1108/ijrdm-03-2013-0062.

Mostafa, M.M. (2009), "Benchmarking the US specialty retailers and food consumer stores using data envelopment analysis", *International Journal of Retail and Distribution Management*, Vol. 37 No. 8, pp. 661-679, doi: 10.1108/09590550910966178.

Nasseri, S.H., Ebrahimnejad, A. and Gholami, O. (2018), "Fuzzy stochastic data envelopment analysis with undesirable outputs and its application to banking industry", *International Journal of Fuzzy Systems*, Vol. 20 No. 2, pp. 534-548, doi: 10.1007/s40815-017-0367-1.

Nyhan, R.C. and Martin, L.L. (1999), "Comparative performance measurement: a primer on data envelopment analysis", *Public Productivity and Management Review*, Vol. 22 No. 3, pp. 348-364, doi: 10.2307/3380708.

Omrani, H., Shafaat, K. and Emrouznejad, A. (2018), "An integrated fuzzy clustering cooperative game data envelopment analysis model with application in hospital efficiency", *Expert Systems with Applications*, Vol. 114, pp. 615-628, doi: 10.1016/j.eswa.2018.07.074.

Pantano, E. and Dennis, C. (2019), "Store buildings as tourist attractions: mining retail meaning of store building pictures through a machine learning approach", *Journal of Retailing and Consumer Services*, Vol. 51, pp. 304-310, doi: 10.1016/j.jretconser.2019.06.018.

Perrigot, R. and Barros, C.P. (2008), "Technical efficiency of French retailers", *Journal of Retailing and Consumer Services*, Vol. 15 No. 4, pp. 296-305, doi: 10.1016/j.jretconser.2007.06.003.

Pestana Barros, C. and Alves, C.A. (2003), "Hypermarket retail store efficiency in Portugal", *International Journal of Retail and Distribution Management*, Vol. 31 No. 11, pp. 549-560, doi: 10.1108/09590550310503285.

Rao, D.S.P., O'Donnell, J. and Battese, G.E. (2003), "Meta-Frontier functions for the study of inter-regional productivity differences", Centre for Efficiency and Productivity Analysis, School of Economics University of Queensland, St Lucia, Working Paper No. 1.

Robinson, D.T. and Caradima, B. (2023), "A multi-scale suitability analysis of home-improvement retail-store site selection for Ontario, Canada", *International Regional Science Review*, Vol. 46 No. 1, pp. 69-97, doi: 10.1177/01600176221092483.

Rouyendegh, B.D., Oztekin, A., Ekong, J. and Dag, A. (2019), "Measuring the efficiency of hospitals: a fully-ranking DEA–FAHP approach", *Annals of Operations Research*, Vol. 278 Nos 1-2, pp. 361-378, doi: 10.1007/s10479-016-2330-1.

Samoilenko, S. and Osei-Bryson, K.M. (2008), "Increasing the discriminatory power of DEA in the presence of the sample heterogeneity with cluster analysis and decision trees", *Expert Systems with Applications*, Vol. 34 No. 2, pp. 1568-1581, doi: 10.1016/j.eswa.2007.01.039.

Samoilenko, S. and Osei-Bryson, K.M. (2010), "Determining sources of relative inefficiency in heterogeneous samples: methodology using Cluster Analysis, DEA and Neural Networks", *European Journal of Operational Research*, Vol. 206 No. 2, pp. 479-487, doi: 10.1016/j.ejor.2010.02.017.

Samoilenko, S. and Osei-Bryson, K.M. (2019), "A data analytic benchmarking methodology for discovering common causal structures that describe context-diverse heterogeneous groups", *Expert Systems with Applications*, Vol. 117, pp. 330-344, doi: 10.1016/j.eswa.2018.09.054.

Sellers-Rubio, R. and Mas-Ruiz, F. (2006), "Economic efficiency in supermarkets: evidences in Spain", *International Journal of Retail and Distribution Management*, Vol. 34 No. 2, pp. 155-171, doi: 10.1108/09590550610649803.

Sharma, M.J. and Yu, S.J. (2009), "Performance based stratification and clustering for benchmarking of container terminals", *Expert Systems with Applications*, Vol. 36 No. 3, pp. 5016-5022, doi: 10.1016/j.eswa.2008.06.010.

Statista (2022), "Retail market worldwide – statistics and facts", available at: https://www.statista.com/topics/5922/retail-market-worldwide/#topicHeader__wrapper

Tasoulis, S., Pavlidis, N.G. and Roos, T. (2020), "Nonlinear dimensionality reduction for clustering", *Pattern Recognition*, Vol. 107, 107508, doi: 10.1016/j.patcog.2020.107508.

Tavana, M., Ebrahimnejad, A., Santos-Arteaga, F.J., Mansourzadeh, S.M. and Matin, R.K. (2018), "A hybrid DEA-MOLP model for public school assessment and closure decision in the City of Philadelphia", *Socio-Economic Planning Sciences*, Vol. 61, pp. 70-89, doi: 10.1016/j.seps.2016.09.003.

Teng, H.S. (2014), "Qualitative productivity analysis: does a non-financial measurement model exist?", *International Journal of Productivity and Performance Management*, Vol. 63 No. 2, pp. 250-256, doi: 10.1108/ijppm-03-2013-0034.

The Business Research Company (2021), "Global market report", available at: https://www.thebusinessresearchcompany.com/report/retail-market

Tsionas, M.G. (2023), "Clustering and meta-envelopment in data envelopment analysis", *European Journal of Operational Research*, Vol. 304 No. 2, pp. 763-778, doi: 10.1016/j.ejor.2022.04.015.

Vaz, C.B., Camanho, A.S. and Guimarães, R.C. (2010), "The assessment of retailing efficiency using network data envelopment analysis", *Annals of Operations Research*, Vol. 173 No. 1, pp. 5-24, doi: 10.1007/s10479-008-0397-z.

Villanueva, B.S., Gibert, K. and Sànchez-Marrè, M. (2013), "Post-processing the class panel graphs: toward understandable patterns from data", in Gibert, K., Botti, V. and Reig-Bolano, R. (Eds), *Artificial Intelligence Research and Development*, IOS Press, pp. 215-224.

Vyt, D. and Cliquet, G. (2017), "Towards a fairer manager performance measure: a DEA application in the retail industry", *International Review of Retail, Distribution and Consumer Research*, Vol. 27 No. 5, pp. 450-467, doi: 10.1080/09593969.2017.1383293.

Yang, F.C. (2020), "Application of centralised DEA in an automobile parts retail network in Taiwan", *International Journal of Retail and Distribution Management*, Vol. 48 No. 7, pp. 667-686, doi: 10.1108/ijrdm-02-2019-0061.

Yu, M.M. and Chen, L.H. (2020), "Evaluation of efficiency and technological bias of tourist hotels by a meta-frontier DEA model", *Journal of the Operational Research Society*, Vol. 71 No. 5, pp. 718-732, doi: 10.1080/01605682.2019.1578625.

Zarrin, M., Schoenfelder, J. and Brunner, J.O. (2022), "Homogeneity and best practice analyses in hospital performance management: an analytical framework", *Health Care Management Science*, Vol. 25 No. 3, pp. 406-425, doi: 10.1007/s10729-022-09590-8.

Zatonatska, T., Wołowiec, T., Dluhopolskyi, O., Podskrebko, O. and Maksymchuk, O. (2022), "Using data science tools in E-commerce: client's advertising campaigns vs Sales of enterprise products", *"International Scientific-Practical Conference" Information Technology for Education, Science and Technics*, Cham, Springer Nature Switzerland, pp. 346-359.

Zhou, W. and Xu, Z. (2020), "An overview of the fuzzy data envelopment analysis research and its successful applications", *International Journal of Fuzzy Systems*, Vol. 22 No. 4, pp. 1037-1055, doi: 10.1007/s40815-020-00853-6.

Zhu, J. (2022), "DEA under big data: data enabled analytics and network data envelopment analysis", *Annals of Operations Research*, Vol. 309 No. 2, pp. 761-783, doi: 10.1007/s10479-020-03668-8.

About the authors

Nicola Castellano (Ph.D.) is an Associate Professor of management accounting at the University of Pisa (Italy), Department of Economics and Management. His main fields of research are management control and performance measurement systems; business analytics; financial disclosure; innovation and performance. Vice-Dean for international relations. Coordinator of the bachelor and the master program in Logistics Management; scientific coordinator of the MBA in Auditing, Finance and Control. Nicola Castellano is the corresponding author and can be contacted at: nicola.castellano@unipi.it

Roberto Del Gobbo (Ph.D.) is a Researcher of management accounting at the University of Macerata (Italy), Department of Economics and Law. His main research interests are corporate performance management and business intelligence, big data and analytics for management control and marketing, and digitalization in corporate governance and control systems. Member of Scientific Committee, MBA in Marketing and Business Management, University of Macerata. Expert management consultant in performance measurement for the manufacturing and retail industry.

Lorenzo Leto is a Ph.D. Student at the University of Pisa (Italy), Department of Economics and Management. His main fields of research are the implementation and development of management control systems and performance measurement systems in small and medium companies; management accounting change and organizational change.

For instructions on how to order reprints of this article, please visit our website:
www.emeraldgrouppublishing.com/licensing/reprints.htm
Or contact us for further details: **permissions@emeraldinsight.com**